

Проектная работа

на тему:

Корпусная лингвистика (корпуса языков, параллельные корпуса). Понятие корпуса, разметки. Виды корпусов. Требования к корпусам. Обзор сетевых ресурсов по корпусной лингвистике. Работа с корпусами. Работа с Национальным корпусом русского языка (НКРЯ). Образовательный портал Национального корпуса русского языка.

Подготовила студентка ВГУ

1 курс

Направление «Лингвистика»

Профиль «ТиМПИЯиК»

Ильичева Светлана Сергеевна

Теория. Обзор интернет ресурсов по корпусной лингвистике.

Корпус языка и наука, которая с этим связана, корпусная лингвистика, — это такая тема, область, которая очень стремительно ворвалась в жизнь лингвистов примерно в самом конце XX — начале XXI века. Если мы хотим назвать такую область лингвистики, которая по определению является суперсовременной, то первое, что приходит в голову, — это как раз лингвистика корпусов.

Корпусная лингвистика — раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов с использованием компьютерных технологий.

Корпус - собрание уникальных, не повторяющихся текстов в электронном, компьютерном, машиночитаемом виде, доступных для поиска, обработки, снабженных разметкой, упрощающей этот поиск, и репрезентативных для данного языка в целом или для какой-то части этого языка.

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. Именно это позволяет использовать параллельный корпус как инструмент исследования.

Корпус языка — это электронное собрание текстов на данном языке.

Национальный корпус русского языка

<http://ruscorpora.ru>

Британский национальный корпус

<http://sara.natcorp.ox.ac.uk/>

Венгерский национальный корпус

<http://corpus.nytud.hu/mnsz/>

Национальный корпус словенского языка

<http://www.fida.net/eng/>

Польский национальный корпус

<http://korpus.ia.uni.lodz.pl/>

Словацкий национальный корпус

<http://korpus.juls.savba.sk>

Чешский национальный корпус

<http://ucnk.ff.cuni.cz>

Таблица 1 Классификация корпусов

Признак	Типы корпусов
Тип данных	Письменные Речевые Смешанные
Язык текстов	Русский Английский и т.д.
«Параллельность»	Одноязычные Двуязычные Многоязычные
«Литературность», Специфичность	Литературные Диалектные Разговорные Терминологические Смешанные
Жанр	Литературные Фольклорные Драматургические Публицистические
Доступность	Свободно доступные Коммерческие Закрытые
Назначение	Исследовательские Иллюстративные
Динамичность	Динамические (мониторные) Статические
Разметка	Размеченные Неразмеченные
Характер разметки	Морфологические Синтаксические Семантические Просодические и т.д.
Объем текстов	Полнотекстовые «Фрагментнотекстовые»
Хронологический аспект	Синхронические Диахронические
«Общность»	Общие Одного писателя
Структура	Центральные и архивные Ядерные и периферийные

Основные требования к корпусам:

1. Репрезентативность - статистически достоверное представление языка или его части, достигаемое за счет необходимого объема и жанрового разнообразия текстов;
2. Разметка - приписывание текстам и их компонентам специальных меток (тэгов).

Обзор сетевых ресурсов:

Linguee - ресурс, использующий имеющиеся в интернете уже переведенные тексты в качестве основы для поиска переводов запросов пользователей. По запросу пользователь получает список соответствующих друг другу законченных предложений на двух языках, которые содержат искомое слово или выражение.

<https://www.linguee.ru/>

Google Books Ngram Viewer — поисковый онлайн-сервис компании Google, позволяющий строить графики частотности языковых единиц на основе огромного количества печатных источников, опубликованных с 16 века и собранных в сервис Google Books. В 2016 году возможен поиск по массивам на американском английском, британском английском, французском, немецком, испанском, итальянском, русском, иврите и упрощенном китайском языках.

<https://books.google.com/ngrams>

НКРЯ <http://ruscorpora.ru/>

Национальный корпус русского языка представляет русский язык в наиболее полном виде: во всём многообразии жанров, стилей, территориальных и социальных вариантов и содержит все типы письменных и устных текстов, представленных в русском языке. В Корпусе собраны художественные тексты разных жанров от Фонвизина до Улицкой, поэзия с конца 18 века, публицистика XX-XXI веков (особенно широко представлена публицистика последних 40 лет), научная литература всех направлений (точные, естественные и гуманитарные науки), официально-деловые тексты: заявления, служебные записки, инструкции, тексты бытовых жанров: мемуары, дневниковые записи, личная переписка, фрагменты интернет-чатов, записи устной разговорной речи, а также записи устной речи из фильмов, диалектные тексты и др.

Образовательный портал Национального корпуса русского языка <https://studiorum-ruscorpora.ru/>

Что Вы найдете на этом сайте?

- В разделе «[Помощь начинающему пользователю](#)» — учебно-методические разработки, которые помогут преподавателям и учителям начать использовать Корпус.
- В разделе «[Методические разработки](#)» — методические материалы, использующие Корпус: упражнения, задания, методики организации самостоятельной работы в Корпусе, а также статьи и монографии на эти темы.
- В разделе «[Корпусная лингвистика сегодня](#)» — информацию о других ресурсах, которые могут быть полезны преподавателям и учителям, интересующимся корпусами.
- В разделе «[Корпусные методы в русистике](#)» — статьи, монографии, авторефераты диссертаций, доклады и тезисы по русскому языку, в которых используются данные Корпуса.
- В разделе «[Корпусное преподавание на Западе](#)» — рефераты англоязычных статей по методике корпусного преподавания.
- В разделе «[Семинар в компьютерном классе](#)» — методические разработки, которые помогут провести занятие с использованием корпуса в компьютерном классе.
- В разделе «[Вопросы к корпусу](#)» — небольшие задания и вопросы, на которые можно получить ответ с помощью корпуса.

- В разделе «Своевольные смыслы: опыт микроисторического исследования лексики 19–21 веков» — материалы корпусного исследования истории значения слов.
- В разделе «Новые издания» — информацию о книгах, написанных с использованием корпуса.
- В разделе «Проблемы русской стилистики по данным НКРЯ» — исследования современного состояния стилистических вариантов в русской речи.

Практика.

Работа в НКРЯ

Начинать работу с Корпусом следует со страницы поиска, куда удобно войти по ссылке «поиск в корпусе» – <http://ruscorpora.ru/search-main.html>

1. Поиск слова, словосочетания или предложения в фиксированной форме

Самый простой вид поиска – это поиск слова, словосочетания или предложения в фиксированной форме. Его удобно использовать для поиска неизменяемых слов, например, наречий, несклоняемых существительных, деепричастий, предлогов, фразеологизмов, а также для поиска точных цитат, включенных в другие тексты. Этим же способом в корпусе можно искать слова или предложения, используемые в русских текстах в латинской записи, такие как IPO или Cogito ergo sum. Кроме того, поиск точных форм следует использовать, чтобы найти определенную морфологическую форму при существовании вариативности, например, при сравнении количества ненормативных форм с нормативными (мерю и меряю, в пальто и в пальте) или для анализа статистики использования вариативных форм (брызгает и брызжет).

Для поиска фиксированных форм предназначена самая верхняя строчка страницы поиска «Поиск точных форм», см. рис 1. Вписав нужную единицу, надо нажать слово «искать» непосредственно под этой строчкой.

Чтобы перейти к поиску следующей единицы, удобно очистить строку поиска при помощи расположенной под ней команды «очистить».

Рис. 1. Запрос для поиска цитаты из песни «Подмосковные вечера»

2. Поиск слова во всех возможных формах

Для поиска изменяемых слов предназначен «Лексико-грамматический поиск». Вписав в его верхнюю строчку слово в начальной форме, мы получим примеры этого слова во всех его формах. Вписав нужную единицу, надо нажать слово «искать» непосредственно под этой строчкой (рис. 2).

Лексико-грамматический поиск ?

Рис. 2. Запрос для поиска глагола представлять во всех формах

3. Поиск слова в нескольких возможных формах

Чтобы найти слово в нескольких определенных формах, например, глагол в формах повелительного наклонения, надо зайти по ссылке «грамматические признаки», в появившейся таблице отметить нужные признаки и затем нажать команду «ОК» (рис. 3).

Лексико-грамматический поиск ?

Часть речи	Падеж	Наклонение / Форма	Степень / Краткость
<input type="checkbox"/> существительное	<input type="checkbox"/> именительный	<input type="checkbox"/> изъявительное	<input type="checkbox"/> сравнительная
<input type="checkbox"/> прилагательное	<input type="checkbox"/> звательный*	<input checked="" type="checkbox"/> повелительное	<input type="checkbox"/> сравнительная 2*
<input type="checkbox"/> числительное	<input type="checkbox"/> родительный	<input checked="" type="checkbox"/> повелительное 2	<input type="checkbox"/> превосходная
<input type="checkbox"/> числ-прил	<input type="checkbox"/> родительный 2	<input type="checkbox"/> инфинитив	<input type="checkbox"/> полная форма
<input type="checkbox"/> глагол	<input type="checkbox"/> дательный	<input type="checkbox"/> причастие	<input type="checkbox"/> краткая форма
<input type="checkbox"/> наречие	<input type="checkbox"/> винительный	<input type="checkbox"/> деепричастие	

Рис. 3.1. Запрос для поиска глагола родиться в формах повелительного наклонения
Можно не называть определенную лексему, а осуществить поиск только по грамматическим признакам, чтобы найти все слова с данным набором признаков.

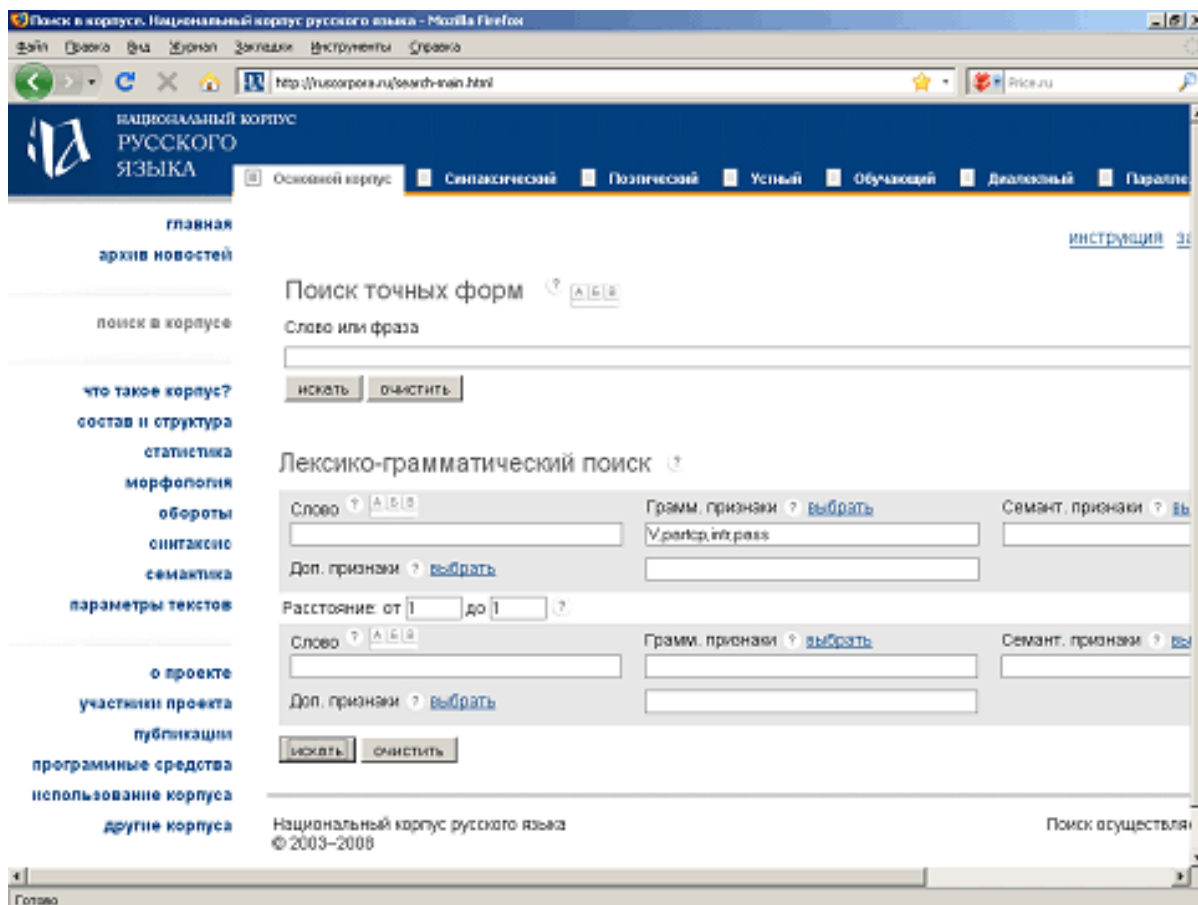


Рис. 3.2. Запрос для поиска всех страдательных причастий, образованных от непереходных глаголов

4. Использование поиска сегмента (морфемы или части основы)

В Корпусе можно искать лексемы по какой-то их части. Например, все слова, которые заканчиваются на –очка. В результате такого запроса можно получить более сорока тысяч примеров существительных на –очка. Чтобы осуществить запрос в месте обрыва сегмента (перед или после него) надо поставить звездочку. См. рис. 4.

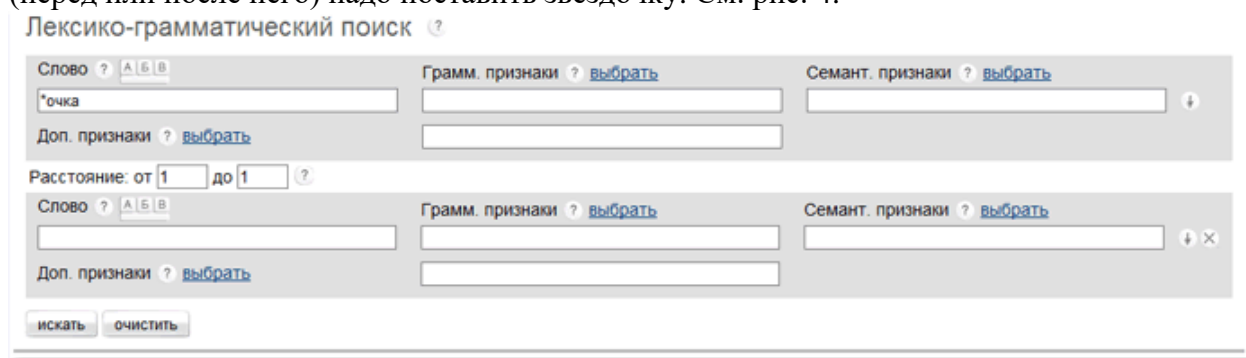


Рис. 4. Запрос для поиска слов, оканчивающихся на –очка

5. Использование поиска слова с определенной семантической характеристикой

Помимо грамматического, в Корпусе есть семантический поиск.

Семантический поиск начинается с команды «семант. признаки. выбрать». Затем в появившемся окне надо выбрать нужную часть речи и в во втором появившемся окне отметить нужные признаки, а затем нажать команду «ОК» (рис. 5).

Каузация

☐ каузативные глаголы

☐ некаузативные глаголы

Служебные глаголы

☒ фазовые

☐ служебные каузативные

Словообразование

☐ приставочные глаголы


☐ семельфактивы

☐ вторичные имперфективы
(-ива-, -ва-, -а-)

Рис. 5. Запрос для семантического поиска: фазовые глаголы

6. Использование поиска нескольких слов

Корпус предоставляет возможность поиска словосочетаний из нескольких слов, где у каждого слова может быть задана грамматическая характеристика. Одно из слов, часть слов или все слова могут быть не конкретными словами, а любым словом с заданной грамматической или семантической характеристикой.

Чтобы появилась возможность добавить к искомым словам еще одно искомое слово, то есть увеличить количество строчек для запроса, следует нажать кнопку  после поля «семант. признаки». Расположенная рядом кнопка «x» убирает лишние строки. См. рис. 6.1.

Лексико-грамматический поиск ?

Слово ? A B V	Грамм. признаки ? выбрать	Семант. признаки ? выбрать
<input type="text" value="каждый"/>	<input type="text"/>	<input type="text"/>
Доп. признаки ? выбрать	<input type="text"/>	
Расстояние: от <input type="text" value="1"/> до <input type="text" value="1"/> ?		
Слово ? A B V	Грамм. признаки ? выбрать	Семант. признаки ? выбрать
<input type="text" value="охотник"/>	<input type="text"/>	<input type="text"/>
Доп. признаки ? выбрать	<input type="text"/>	
Расстояние: от <input type="text" value="1"/> до <input type="text" value="1"/> ?		
Слово ? A B V	Грамм. признаки ? выбрать	Семант. признаки ? выбрать
<input type="text" value="желает"/>	<input type="text"/>	<input type="text"/>
Доп. признаки ? выбрать	<input type="text"/>	
Расстояние: от <input type="text" value="1"/> до <input type="text" value="1"/> ?		
Слово ? A B V	Грамм. признаки ? выбрать	Семант. признаки ? выбрать
<input type="text" value="знать"/>	<input type="text"/>	<input type="text"/>
Доп. признаки ? выбрать	<input type="text"/>	
<input type="button" value="искать"/> <input type="button" value="очистить"/>		

Рис. 6.1. Запрос для поиска длинной фразы

Также возможно найти слова, не обязательно находящиеся рядом. Это может понадобиться, например, если второе слово искомого словосочетания существительное и не хочется терять примеры, в которых ему предшествует любое прилагательное. То есть если нужно найти оборот как щенок, не теряя при этом сочетаний как слепой щенок или как годовалый щенок. Или, допустим, нужно найти идиоматическое выражение, в котором устойчивыми членами являются два слова (например, будить зверя), а между ними могут помещаться разные другие слова (например, будить в ком-либо зверя или будить какого-либо зверя и др.) (рис. 6.2).

Лексико-грамматический поиск ?

Слово ? A B V	Грамм. признаки ? выбрать	Семант. признаки ? выбрать
<input type="text" value="*будить"/>	<input type="text"/>	<input type="text"/>
Доп. признаки ? выбрать	<input type="text"/>	
Расстояние: от <input type="text" value="1"/> до <input type="text" value="5"/> ?		
Слово ? A B V	Грамм. признаки ? выбрать	Семант. признаки ? выбрать
<input type="text" value="зверь"/>	<input type="text"/>	<input type="text"/>
Доп. признаки ? выбрать	<input type="text"/>	
<input type="button" value="искать"/> <input type="button" value="очистить"/>		

Рис. 6.2. Запрос для поиска идиомы «будить (до 5 любых слов) зверя»

Чтобы найти два слова, находящиеся рядом, но не обязательно расположенные подряд друг за другом, следует задать ограничение на расстояние между нужными словами в поле «Расстояние» (рис. 6.2).

Лексико-грамматический поиск ?

Слово ? A B V	Грамм. признаки ? выбрать	Семант. признаки ? выбрать
<input type="text" value="как"/>	<input type="text"/>	<input type="text"/>
Доп. признаки ? выбрать	<input type="text"/>	
Расстояние: от <input type="text" value="1"/> до <input type="text" value="3"/> ?		
Слово ? A B V	Грамм. признаки ? выбрать	Семант. признаки ? выбрать
<input type="text"/>	<input type="text" value="S"/>	<input type="text" value="r.concr & t:animal"/>
Доп. признаки ? выбрать	<input type="text"/>	
<input type="button" value="искать"/> <input type="button" value="очистить"/>		

Запрос для поиска сочетания союза как с названием животного при расстоянии между словами не более трех

7. Использование поиска слова с учетом повторов, заглавной буквы или с учетом ближайших знаков препинания

Воспользовавшись разделом «дополнительные признаки» («доп. признаки»), можно задать поиск слова или определенной грамматической формы, которая повторяется в ближайшем контексте. При помощи опции «дополнительные признаки» можно искать слова, написанные с заглавной буквы.

Также можно искать слова, находящиеся до или после определенного знака препинания.

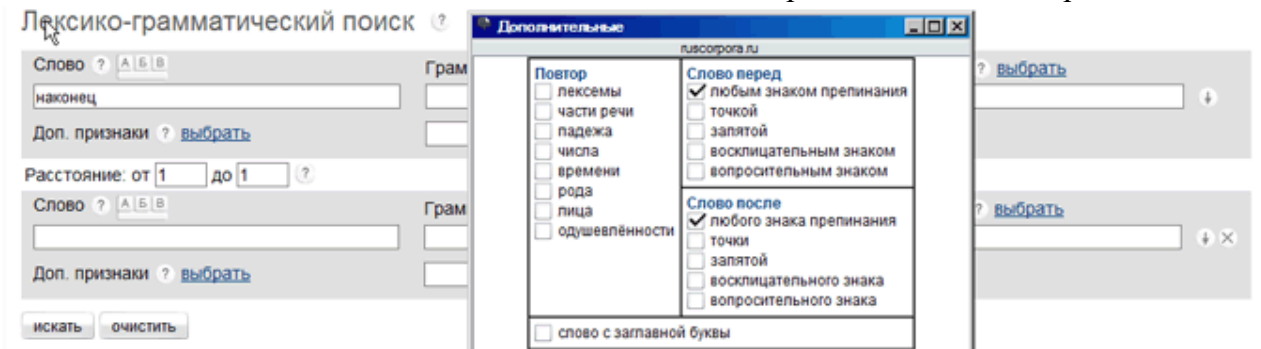


Рис. 7. Запрос для поиска слова наконец, выделенного запятыми

Личный подкорпус

Одной из самых необходимых возможностей, предоставляемых Корпусом, является возможность осуществлять поиск не сразу во всех текстах, а только в тех, которые интересуют пользователя. При этом можно отобрать тексты по самым разным параметрам:

- по произведению или автору;
- по жанру или тематике текстов;
- по времени создания текстов;
- месту и времени описываемых событий.

Советы пользователю:

Чтобы осуществить поиск в текстах по выбору пользователя, надо войти по ссылке «задать подкорпус», расположенной в верхнем правом углу страницы поиска.

Обратите внимание, что при выборе жанров и тематики возможен инвертированный выбор: если нужны все признаки, кроме какого-то одного, то надо пометить «ненужный» признак и нажать расположенную над списком кнопку «инвертировать выбор» (см. рис. 8.2).

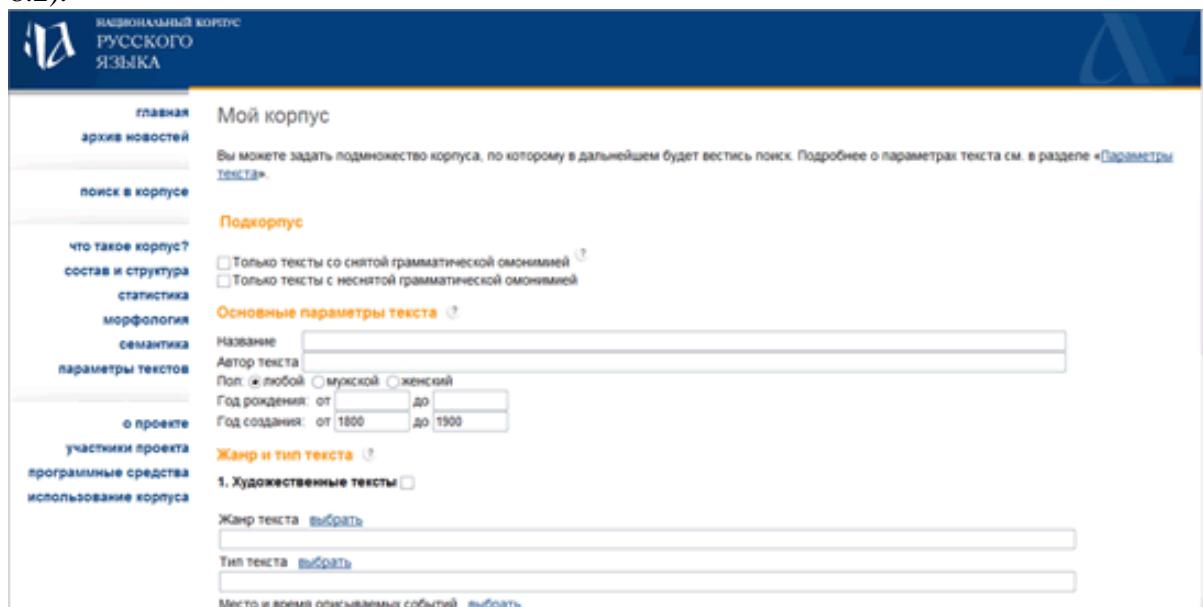


Рис. 8.1. Запрос для создания подкорпуса, ограниченного текстами XIX века

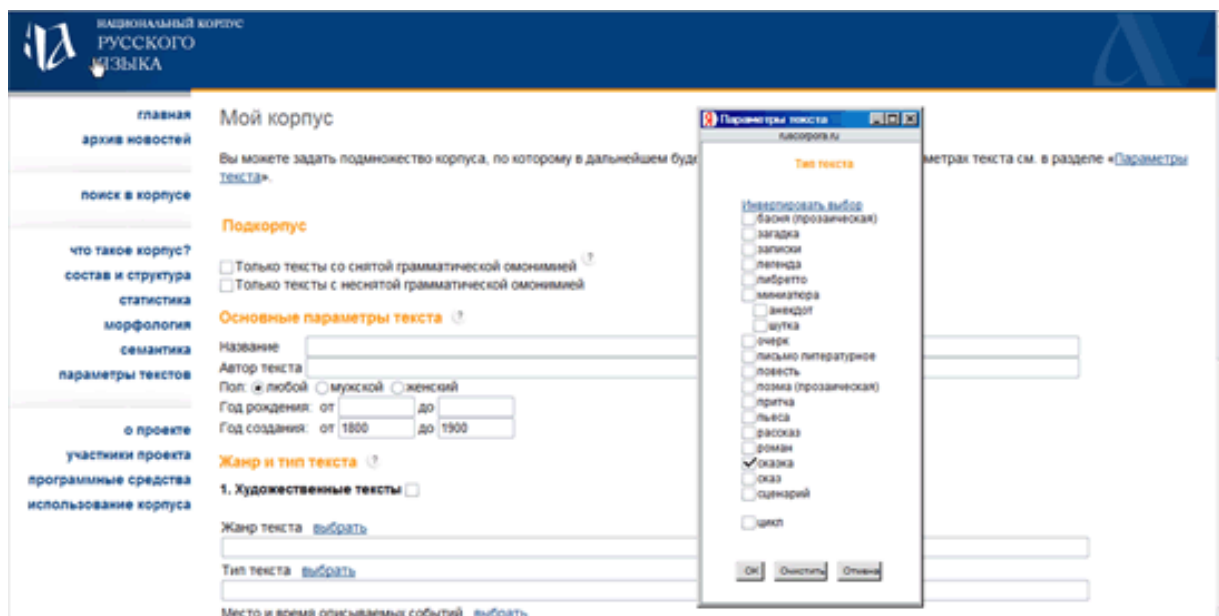


Рис. 8.2. Запрос для создания подкорпуса, ограниченного типом художественного текста

Материалы для закрепления изученного

1. Упражнение:

При помощи найденных в корпусе примеров вставьте в места пропусков глаголы представить или предоставить:

1. Российский Минобраз_____регионам право самим решать, при какой температуре воздуха дети могут не ходить в школу.

2. Когда зал наполнился приглашенными, ведущий открыл встречу и _____слово заместителю губернатора области.

3. Главу своей будущей книги писатель любезно_____для публикации в нашей газете.

4. Почти ста участникам круглого стола будут_____строительные возможности и перспективы Краснодарского края.

5. Здесь не отказывают в социальной помощи никому из тех, кто_____нужные справки.

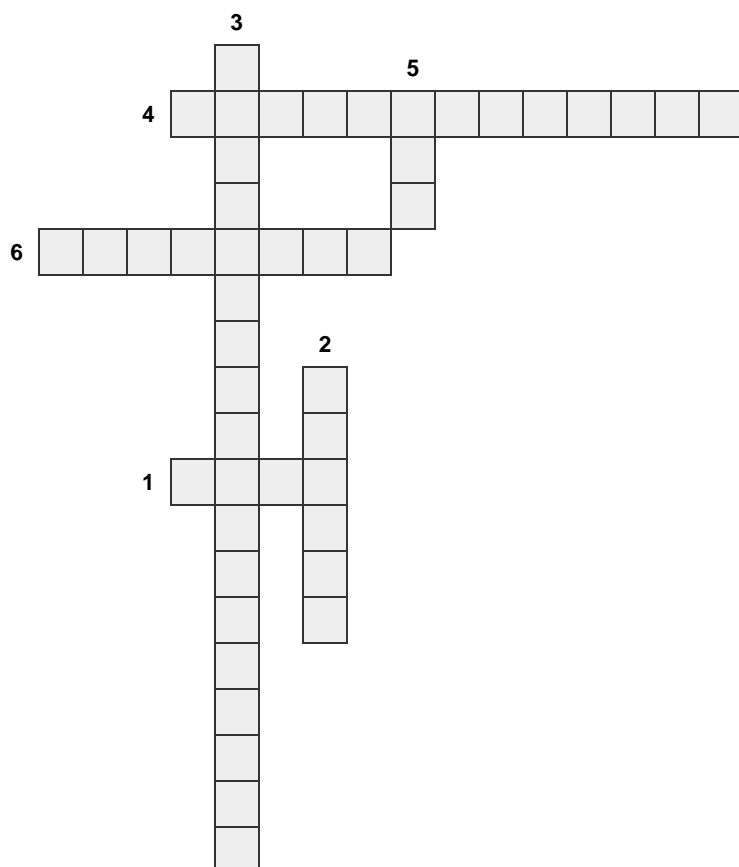
6. Музей в пятый раз подряд безвозмездно_____помещение для торжественной церемонии.

2. Задание: Найдите в корпусе примеры использования слов *большОй* и *бОльший* во всех формах.

3. Тест: <https://ru.surveymonkey.com/r/X7FKCDR>

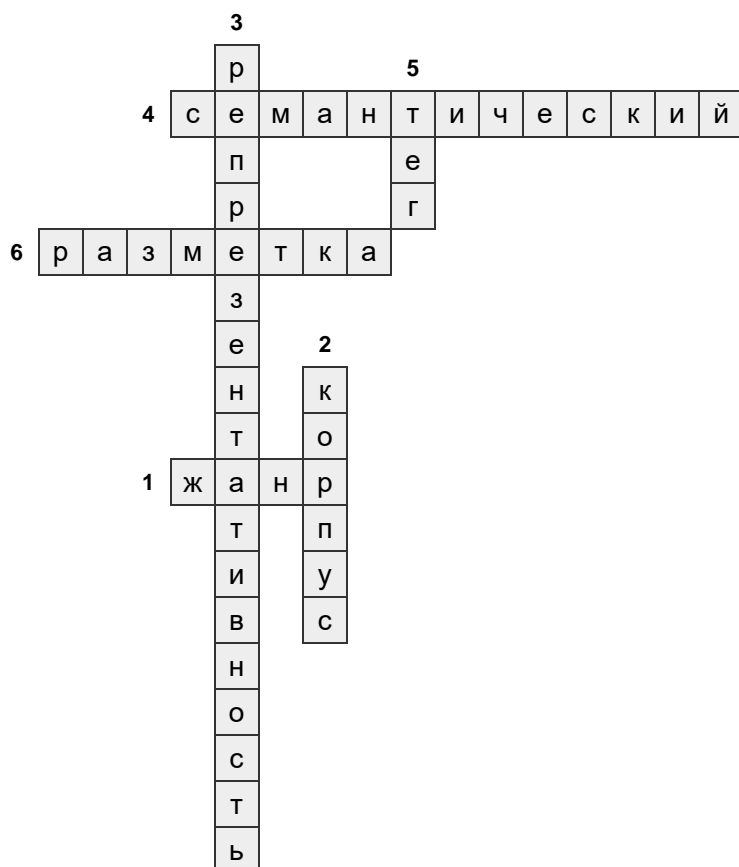
4. Карточки: https://quizlet.com/_5vblsc

5. Кроссворд:



1. Признак, по которому корпуса делятся на: литературные, фольклорные, драматургические, публицистические
2. собрание уникальных, не повторяющихся текстов в электронном, компьютерном, машиночитаемом виде, доступных для поиска, обработки, снабженных разметкой, упрощающей этот поиск, и репрезентативных для данного языка в целом или для какой-то части этого языка.
3. статистически достоверное представление языка или его части, достигаемое за счет необходимого объема и жанрового разнообразия текстов;
4. тип корпуса по характеру разметки
5. специальные метки, приписываемые текстам и их компонентам
6. приписывание текстам и их компонентам специальных меток

Ответы на кроссворд:



Литература:

<https://postnauka.ru/video/7783>

<https://postnauka.ru/longreads/54875>

<http://polit.ru/article/2013/05/02/nkrya>

<https://studiorum-ruscorpora.ru/help/>

<https://www.myfilology.ru/177/korpusy-i-korpusnaya-lingvistika-osnovnye-ponyatiya/>