Федеральное государственное высшее учебное заведение Национальный исследовательский университет «Высшая школа экономики»

Корпус русских учебных текстов

Проект подготовили:
Терехина Лилия
Николаенкова Мария

группа 4

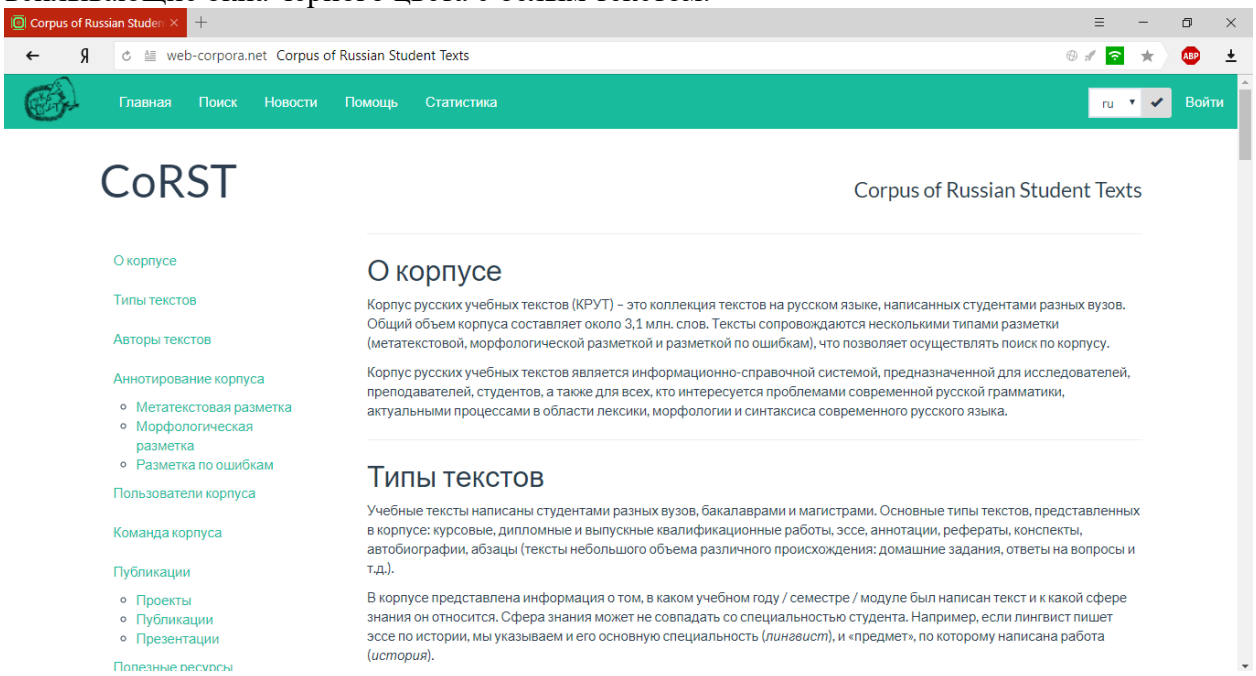
Москва, 2017 г.

Немного о корпусе

В настоящее время существует огромное количество корпусов, которые созданы для выполнения разных задач. Для своего проекта мы выбрали Корпус русских учебных текстов (КРУТ/ Corpus of Russian Student Texts/ CoRST), который включат в себя тексты на русском языке, написанные студентами разных вузов. Он обладает достаточно большим объемом слов, несколькими типами разметки и удобным поиском.
Корпус русских учебных текстов – уникальный лингвистический ресурс для студентов, исследователей, которые опираются на материалы корпуса для выполнения различных проектов, подведения статистик, и для людей, которым интересны природа ошибок, совершаемых носителями русского языка, а также современный язык в целом.

Дизайн

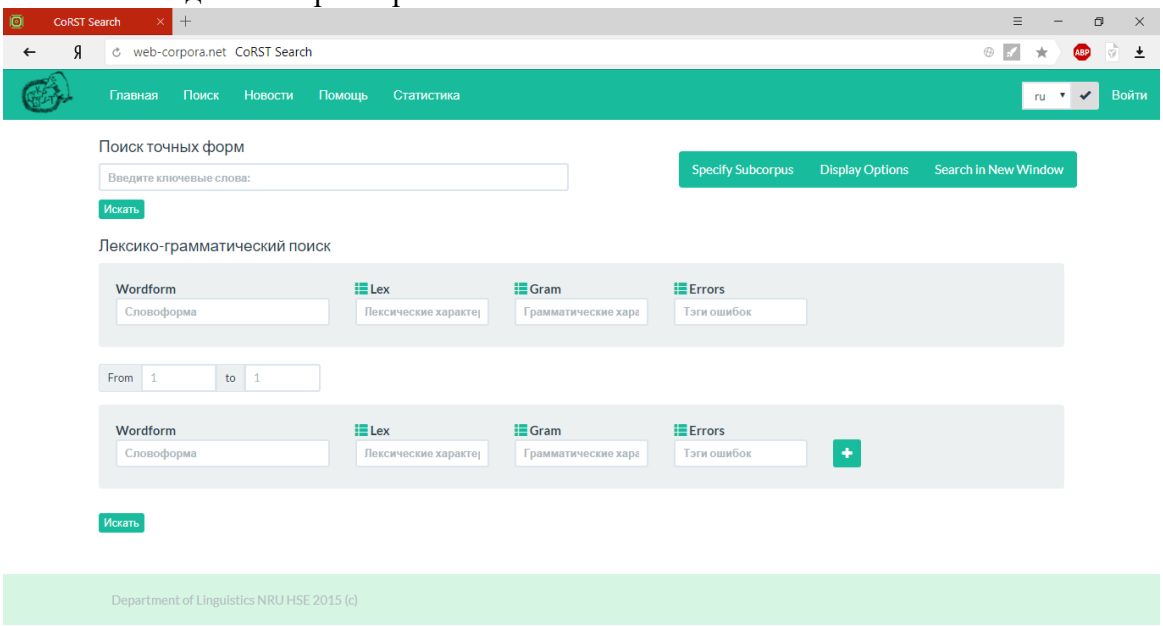
При оформлении корпуса использовались стандартные цвета, шрифты, расположение блоков, которые присутствуют и в других известных корпусах. Основные цвета сайта: белый и зеленый, цвет текста: черный. Для выделения используются желтый и серые цвета и полужирное начертание. Данное сочетание очень удобно для пользователей, сайт не нагружен лишней информацией, приятен глазу. На некоторых страницах есть всплывающие окна черного цвета с белым текстом.



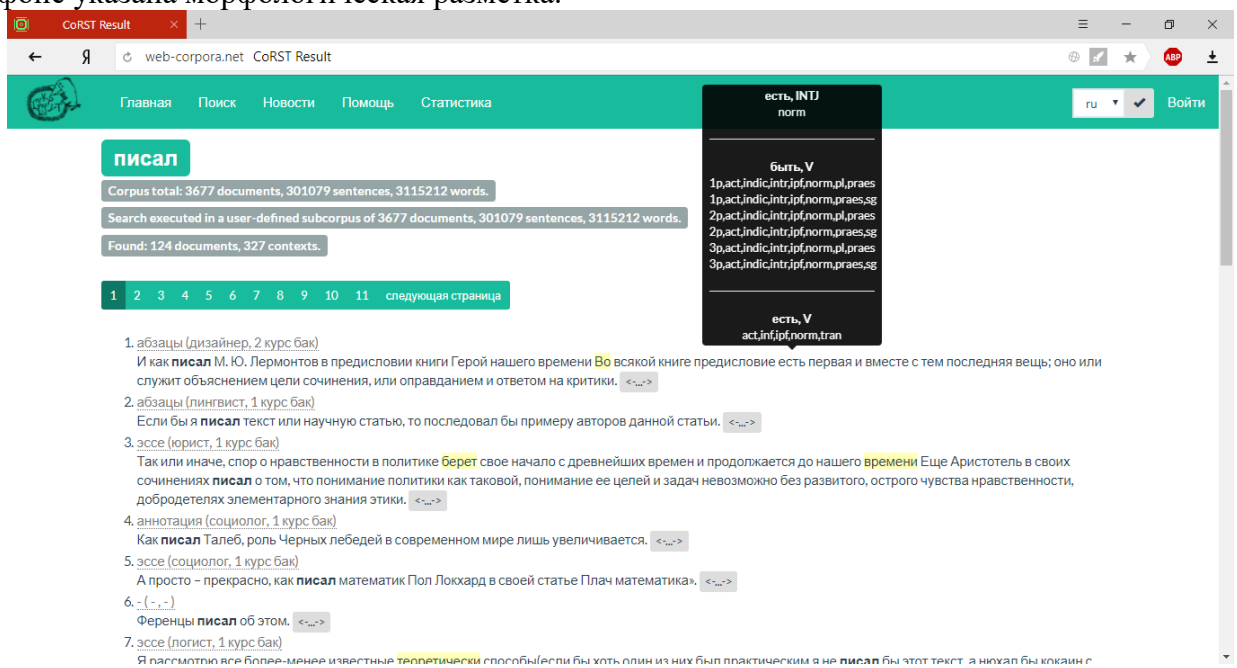
Ресурс глазами новичка

Сайт находится в поисковой системе просто – нужно всего лишь ввести название в поисковую строку, и первым результатом будет наш сайт. Корпус создан на основе webcorpora.net, что объединяет его с остальными известными корпусами. Ссылки на данный ресурс так же можно найти в разделе «Другие корпуса» в НКРЯ и на сайтах с учебными материалами.
Главная страница имеет несколько блоков: в верхней части расположено главное меню, которое не исчезает при переходе на остальные страницы. На нем находятся эмблема корпуса, возврат на главную страницу, переход к поиску, новости, помощь и статистика и выбор языка. Сайт доступен на двух языках (русский и английский).

Слева имеется быстрый доступ к информации, расположенной на главной странице: о создателях, разметке, типах текста и полезные ресурсы.
Кнопка поиска расположена в верхнем меню, достаточно заметна. Ресурс предоставляет поиск точных форм и лексико-грамматический поиск по словоформам.
Страница поиска напоминает поиск в НКРЯ, где можно вводить несколько словоформ, указывая необходимые параметры.



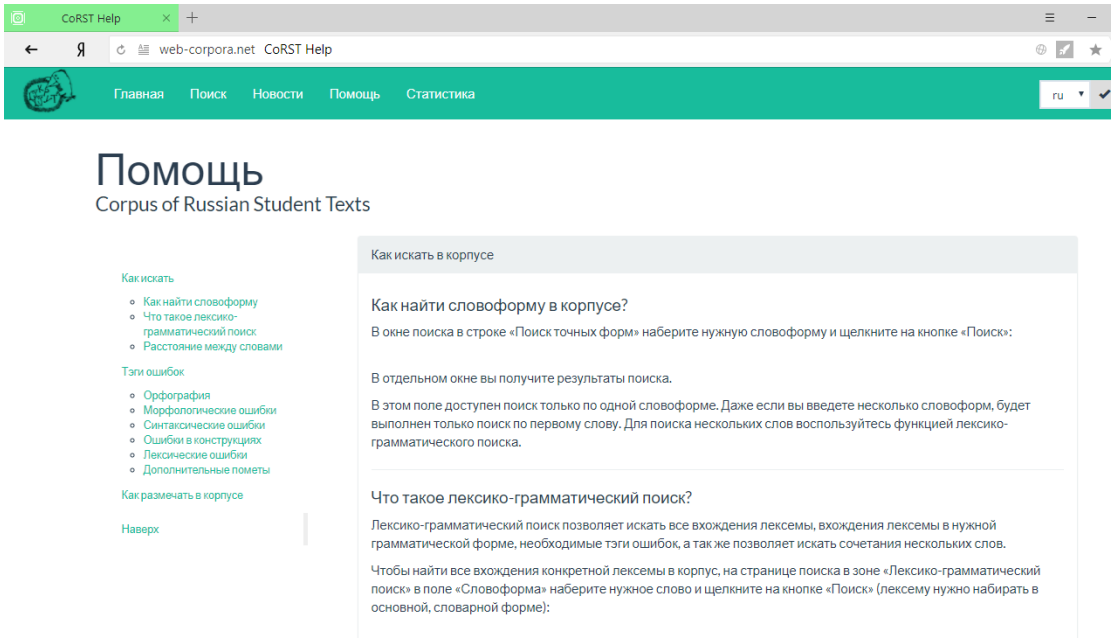
На **странице выдачи результатов** располагается по тридцать примеров, к нужной лемме применяется полужирное начертание, ошибки выделяются желтым цветом. Кнопка перехода на другие страницы расположена как в начале, так и в конце.
Также указана информация, сколько слов и документов включает в себя корпус, количество примеров и документов
Для каждого результата показаны тип текста, специальность автора и его курс, небольшой контекст.
При наведении курсора на слово в поиске появляется окно, где белым текстом на черном фоне указана морфологическая разметка.



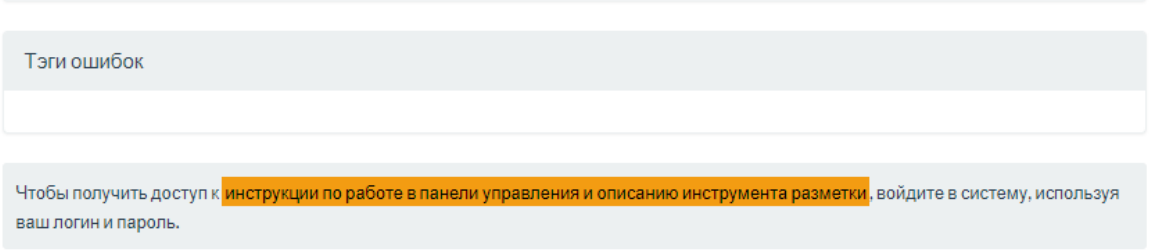
Помощь

При возникновении проблем пользователь может перейти в раздел «Помощь» , щелкнув на кнопку в верхнем меню. Там расположены ответы на самые распространенные вопросы. Присутствуют инструкция по обоим типам поиска (поиск точных форм и лексико-грамматический поиск), однако отсутствуют образцы поиска и скриншоты.
Вопросы, ответы на которые присутствуют в разделе «Помощь»:

- Как найти словоформу в корпусе?
- Что такое лексико-грамматический поиск?
- Расстояние между словами
- Тэги ошибок

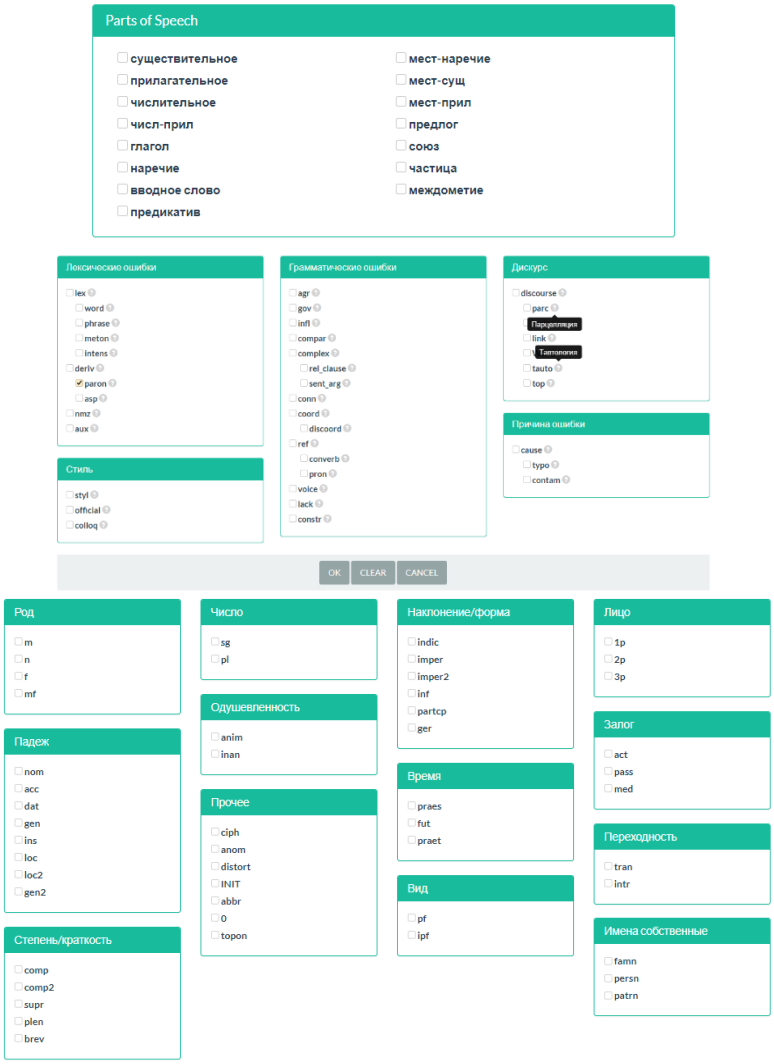


Доступ к разделу «Тэги ошибок» закрыт, для него необходима авторизация на сайте, что значительно затрудняет работу с ресурсом.



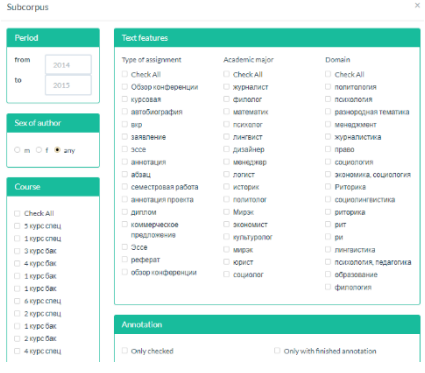
Продвинутый функционал

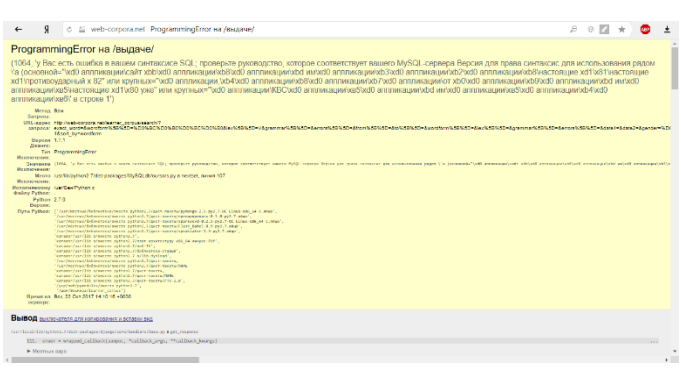
В нашем корпусе можно выполнять поиск по частям речи, по грамматическим признакам, по типу ошибки. Хотим отметить, что при выборе ошибки на странице поиска, при нажатии на «?» появляется черное окно с белым шрифтом, где указано традиционное русское название ошибки.



Корпус русских учебных текстов позволяет совершать поиск с помощью задания подкорпусов, в котором можно указать пол автора, год обучения, специальность, вид работы.

Корпус работает нестабильно, очень часто выдает ошибки в виде странной страницы, что может напугать пользователя.

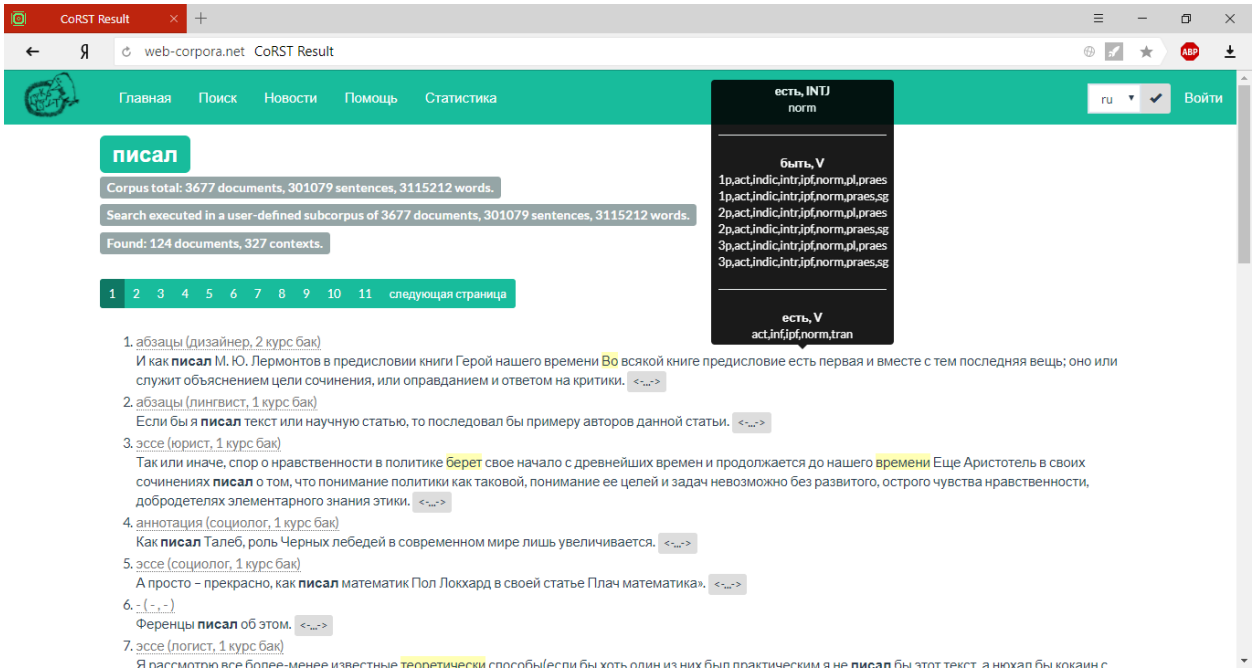




Примеры запросов

Поиск точных форм:

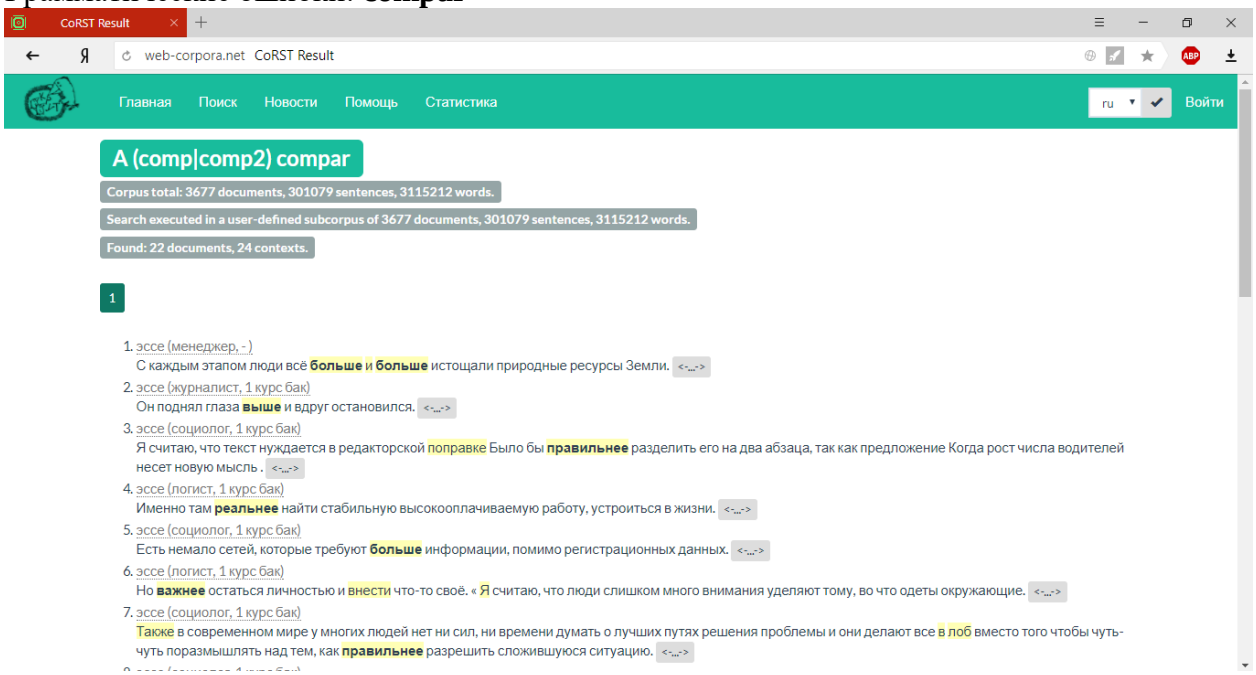
Для примера найдем примеры, в которых присутствует словоформа «писал»:



Лексико-грамматический поиск:

Найдем примеры с ошибкой в употреблении степеней прилагательных:

Запрос: **Lex**: прилагательное; **Gram**: Степень/краткость: comp+comp2; **Errors**: Грамматические ошибки: **compar**

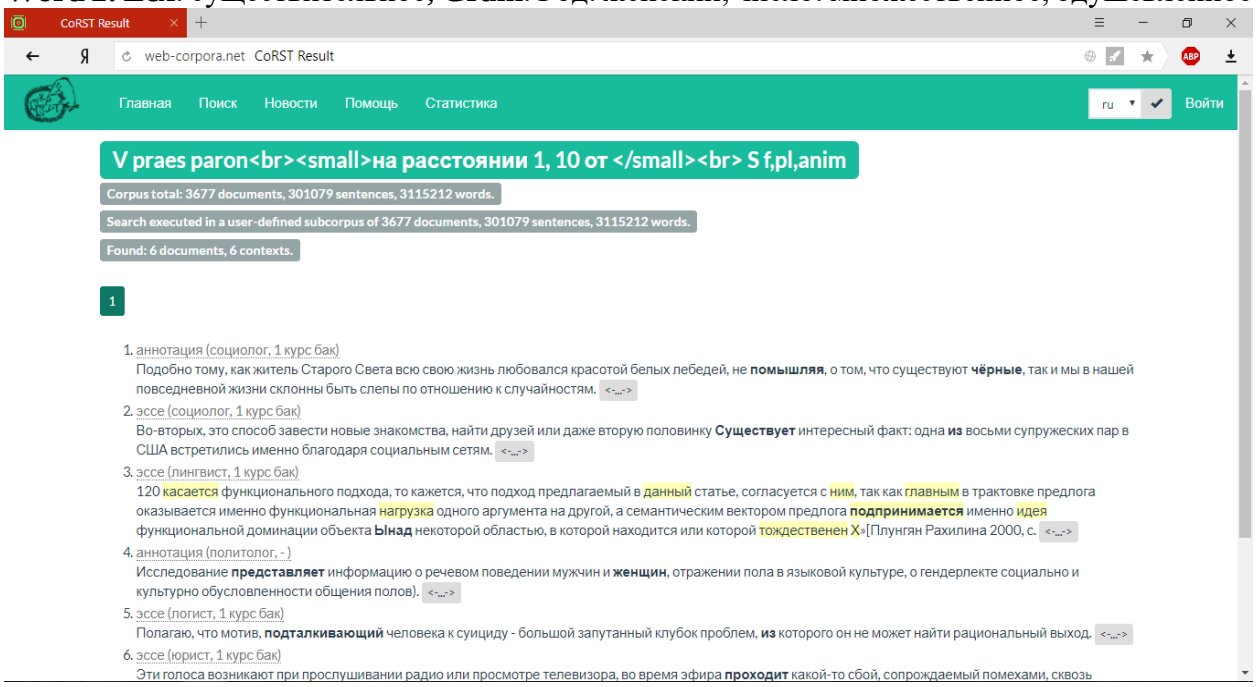


Поиск двух слов:

В корпусе возможно осуществить поиск по нескольким словам на разных расстояниях друг от друга. Зададим запрос для двух слов , одушевленного существительного женского рода множественного числа и глагола настоящего времени, на расстоянии от 1 до 10 с ошибкой в употреблении паронимов:

Word 1: Lex: глагол; **Gram**: Время: настоящее; **Errors**: Лексические ошибки: **paron**

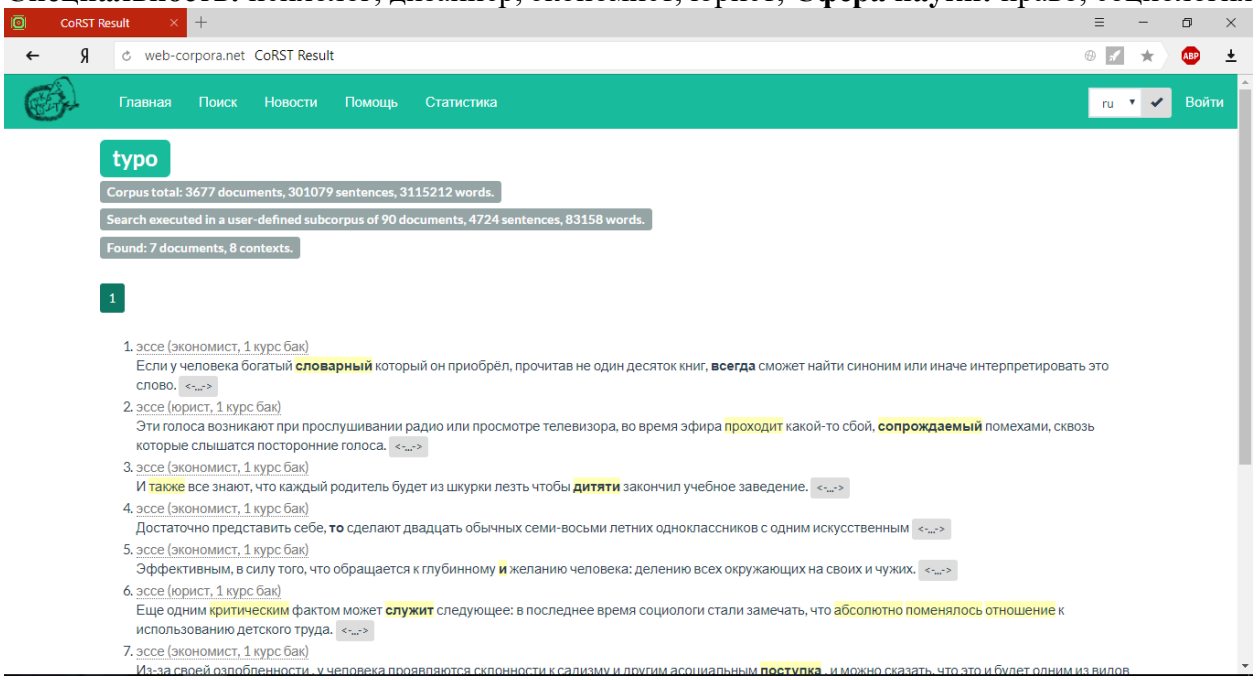
Word 2: Lex: существительное; **Gram**: Род: женский, число: множественное, одушевленное



Поиск в подкорпусе:

Найдем примеры с опечатками среди текстов, авторами которых являются психологи, дизайнеры, экономисты, юристы первого курса бакалавриата и специалитета мужского пола в сферах права и социологии:

Errors: typo; **Пол:** мужской; **Курс:** 1 курс спец, 1 курс бак; **Тип текста:** Выбрать все; **Специальность:** психолог, дизайнер, экономист, юрист; **Сфера науки:** право, социология



Публикации

- *Zevakhina N., Dzhakupova S. Глава «Corpus of Russian student texts: design and prospects» в книге «Материалы 21-й Международной конференции по компьютерной лингвистике "Диалог"» М.: Изд-во РГГУ, 2015. Язык : английский, ПУБЛИКАЦИЯ ПОДГОТОВЛЕНА ПО РЕЗУЛЬТАТАМ ПРОЕКТА: Корпусные исследования границ речевого варьирования: от аграмматизма к норме(2015)*

  В главе содержатся причины создания Корпуса учебных текстов, его основные цели, структура, информация о типах разметки, яркие примеры, перспективы данного ресурса
- *Пужаева С. Ю., Зевахина Н. А., Джакупова С. С. Глава «Контаминация конструкций в речи нестандартных русскоговорящих на материале корпуса русских учебных текстов» В книге «Труды Международной научной конференции "Корпусная лингвистика-2015"» СПб.: Издательство СПбГУ, 2015. ПУБЛИКАЦИЯ ПОДГОТОВЛЕНА ПО РЕЗУЛЬТАТАМ ПРОЕКТА: Корпусные исследования границ речевого варьирования: от аграмматизма к норме (2015).*

  В докладе проводится поиск наиболее весомых причин ошибок, совершаемых студентами, в различных конструкциях. Материалом исследования служат примеры, взятые из Корпуса русских учебных текстов

Разметки

В КРУТ имеется три разметки: метатекстовая, морфологическая, разметка по ошибкам, что, несомненно, является преимуществом. Для сравнения: в НРКЯ имеется пять разметок.
Морфологическая разметка. Это главный тип разметки в текстах, он является основой для поиска и дальнейшего анализа. В корпусе учебных русских текстов используется та же разметка, что и в НРКЯ. Также она осуществляется автоматически с помощью программы MYSYSTEM
Разметка по ошибкам. Многоуровневая система разметки по ошибкам в данный момент проходит апробацию, поэтому результаты иногда выдаются некорректно или вообще не выводятся. Разработчики в скором времени обещают разметить тексты вручную. Именно эта разметка является уникальной и отличает КРУТ от иных лингвистических ресурсов.
Метатекстовая разметка. Она позволяет пользователю ограничить область поиска текстами определенного типа. Является довольно важной составляющей корпуса, поскольку для многих исследований лингвистам необходима информация об авторе текста и о самом тексте для подведения различных статистик.

Недостатки в работе корпуса

- Единственный минус, касающийся поиска: в окне выбора грамматических и лексических параметров используются сокращенные пометы («теги») на основе латинского алфавита, положенного в основу метаязыка грамматических помет без расшифровки на русский язык. Ссылка на список с расшифровкой присутствует на главной странице корпуса. Но это достаточно неудобно, так как приходится вручную искать расшифровку нужного тега.
- К сожалению, Корпус учебных русских текстов работает довольно нестабильно. Возможно, ресурс пока что не справляется со сложными запросами, когда пользователь задает множество параметров, несколько словоформ, внутри подкорпуса.
- Также в корпусе отсутствует возможность сортировать результаты поиска (по алфавиту, по дате создания текста). При необходимости придется делать это

вручную. Более того, выданные данные нельзя скачать ни в каком виде для работы в режиме офлайн.

- Можно отметить и отсутствие примеров поиска, видеоинструкций и неполный перевод некоторых разделов сайта.