

# Корпус русских учебных текстов

Проект подготовили:

Терехина Лилия

Николаенкова Мария

группа 4

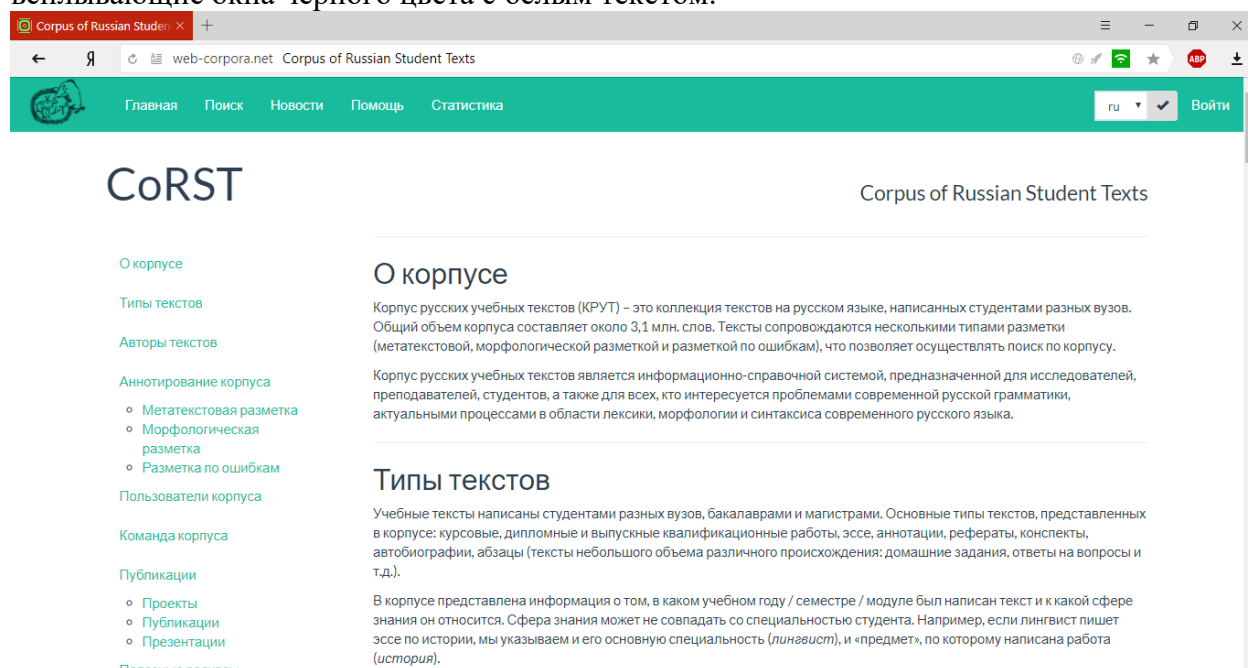
# Немного о корпусе

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 включают 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.

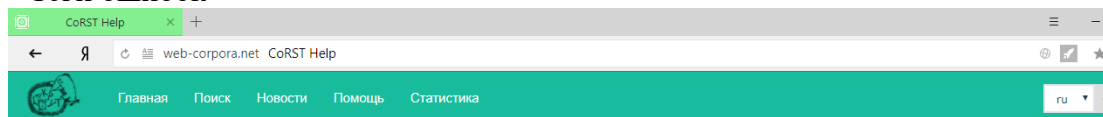
При наведении курсора на слово в поиске появляется окно, где белым текстом на черном фоне указана морфологическая разметка.

# 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



## Помощь

Corpus of Russian Student Texts

### Как искать

- Как найти словоформу
- Что такое лексико-грамматический поиск
- Расстояние между словами

### Тэги ошибок

- Орфография
- Морфологические ошибки
- Синтаксические ошибки
- Ошибки в конструкциях
- Лексические ошибки
- Дополнительные пометы

### Как размечать в корпусе

Наверх

### Как искать в корпусе

#### Как найти словоформу в корпусе?

В окне поиска в строке «Поиск точных форм» наберите нужную словоформу и щелкните на кнопке «Поиск»:

В отдельном окне вы получите результаты поиска.

В этом поле доступен поиск только по одной словоформе. Даже если вы введете несколько словоформ, будет выполнен только поиск по первому слову. Для поиска нескольких слов воспользуйтесь функцией лексико-грамматического поиска.

#### Что такое лексико-грамматический поиск?

Лексико-грамматический поиск позволяет искать все вхождения лексемы, вхождения лексемы в нужной грамматической форме, необходимые тэги ошибок, а так же позволяет искать сочетания нескольких слов.

Чтобы найти все вхождения конкретной лексемы в корпус, на странице поиска в зоне «Лексико-грамматический поиск» в поле «Словоформа» наберите нужное слово и щелкните на кнопке «Поиск» (лексему нужно набирать в основной, словарной форме):

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

### Тэги ошибок

Чтобы получить доступ к инструкции по работе в панели управления и описанию инструмента разметки, войдите в систему, используя ваш логин и пароль.

## 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.

Parts of Speech

☐ существительное
 ☐ местоимение
 ☐ место-название
 ☐ место-сущ
 ☐ место-прил
 ☐ предлог
 ☐ союз
 ☐ частица
 ☐ междометие

☐ прилагательное
 ☐ числительное
 ☐ числ-прил
 ☐ глагол
 ☐ наречие
 ☐ вводное слово
 ☐ предикатив

Лексическая ошибка

☐ lex
 ☐ word
 ☐ phrase
 ☐ meton
 ☐ intens
 ☐ deriv
 ☒ paron
 ☐ asp
 ☐ nrmz
 ☐ aux

Стиль

☐ styl
 ☐ official
 ☐ colloq

Грамматическая ошибка

☐ agr
 ☐ gov
 ☐ infl
 ☐ compar
 ☐ complex
 ☐ rel\_clause
 ☐ sent\_arg
 ☐ comn
 ☐ coord
 ☐ discord
 ☐ ref
 ☐ convert
 ☐ pron
 ☐ voice
 ☐ lack
 ☐ constr

Дискурс

☐ discourse
 ☐ parc
 ☒ Topicmarker
 ☐ link
 ☒ Transition
 ☐ tauto
 ☐ top

Примечание ошибки

☐ cause
 ☐ typo
 ☐ contain

OK

CLEAR

CANCEL

Род

☐ m
 ☐ n
 ☐ f
 ☐ mf

Число

☐ sg
 ☐ pl

Одушевленность

☐ anim
 ☐ inan

Прочее

☐ ciph
 ☐ anom
 ☐ distort
 ☐ INIT
 ☐ abbr
 ☐ 0
 ☐ topon

Наклонение/форма

☐ indic
 ☐ imper
 ☐ imper2
 ☐ inf
 ☐ partcp
 ☐ ger

Время

☐ praes
 ☐ fut
 ☐ praet

Вид

☐ pf
 ☐ ipf

Лицо

☐ 1p
 ☐ 2p
 ☐ 3p

Залог

☐ act
 ☐ pass
 ☐ med

Переходность

☐ tran
 ☐ intr

Степень/градусность

☐ comp
 ☐ comp2
 ☐ supr
 ☐ plen
 ☐ gen2
 ☐ brev

Имена собственные

☐ famn
 ☐ persn
 ☐ patrn

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.

[illegible]

# Примеры запросов

## Поиск точных форм:

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

The screenshot shows the CoRST Result web interface. The search term is 'писал'. The results show a corpus of 3677 documents, 301079 sentences, and 3115212 words. The search was executed in a user-defined subcorpus of 3677 documents, 301079 sentences, and 3115212 words. Found: 124 documents, 327 contexts.

Navigation links: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, следующая страница

1. абзацы (дизайнер, 2 курс бак)  
И как **писал** М. Ю. Лермонтов в предисловии книги Герой нашего времени **Во** всякой книге предисловие есть первая и вместе с тем последняя вещь; оно или служит объяснением цели сочинения, или оправданием и ответом на критику. <...>

2. абзацы (лингвист, 1 курс бак)  
Если бы я **писал** текст или научную статью, то последовал бы примеру авторов данной статьи. <...>

3. эссе (юрист, 1 курс бак)  
Так или иначе, спор о нравственности в политике **берет** свое начало с древнейших времен и продолжается до нашего **времени**. Еще Аристотель в своих сочинениях **писал** о том, что понимание политики как таковой, понимание ее целей и задач невозможно без развитого, острого чувства нравственности, добродетелей элементарного знания этики. <...>

4. аннотация (социолог, 1 курс бак)  
Как **писал** Талеб, роль Черных лебедей в современном мире лишь увеличивается. <...>

5. эссе (социолог, 1 курс бак)  
А просто – прекрасно, как **писал** математик Пол Локхард в своей статье Плач математика. <...>

6. - (-; -)  
Ференцы **писал** об этом. <...>

7. эссе (логист, 1 курс бак)  
Я рассматриваю все более-менее известные **тропические** способы (если бы хоть один из них был практическим я не **писал** бы этот текст, а купил бы кокаин с

## Лексико-грамматический поиск:

Найдем примеры с ошибкой в употреблении степеней прилагательных:

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

Грамматические ошибки: **compar**

The screenshot shows the CoRST Result web interface. The search term is 'A (comp|comp2) compar'. The results show a corpus of 3677 documents, 301079 sentences, and 3115212 words. The search was executed in a user-defined subcorpus of 3677 documents, 301079 sentences, and 3115212 words. Found: 22 documents, 24 contexts.

1

1. эссе (менеджер, -)  
С каждым этапом люди все **больше и больше** истощали природные ресурсы Земли. <...>

2. эссе (журналист, 1 курс бак)  
Он поднял глаза **выше** и вдруг остановился. <...>

3. эссе (социолог, 1 курс бак)  
Я считаю, что текст нуждается в редакторской **поправке**. Было бы **правильнее** разделить его на два абзаца, так как предложение Когда рост числа водителей несет новую мысль. <...>

4. эссе (логист, 1 курс бак)  
Именно там **реальнее** найти стабильную высокооплачиваемую работу, устроиться в жизнь. <...>

5. эссе (социолог, 1 курс бак)  
Есть немало сетей, которые требуют **больше** информации, помимо регистрационных данных. <...>

6. эссе (логист, 1 курс бак)  
Но **важнее** остаться личностью и **внести** что-то своё. «Я считаю, что люди слишком много внимания уделяют тому, во что одеты окружающие. <...>

7. эссе (социолог, 1 курс бак)  
**Также** в современном мире у многих людей нет ни сил, ни времени думать о лучших путях решения проблемы и они делают все **влюб** вместо того чтобы чуть-чуть поразмышлять над тем, как **правильнее** разрешить сложившуюся ситуацию. <...>

## 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

CoRST Result

web-corpora.net CoRST Result

Главная Поиск Новости Помощь Статистика

ru Войти

V praes paron<br><small>на расстоянии 1, 10 от </small><br> S f,pl,anim

Corpus total: 3677 documents, 301079 sentences, 3115212 words.

Search executed in a user-defined subcorpus of 3677 documents, 301079 sentences, 3115212 words.

Found: 6 documents, 6 contexts.

1

1. аннотация (социолог, 1 курс бак)  
Подобно тому, как житель Старого Света всю свою жизнь любовался красотой белых лебедей, не **помышляя**, о том, что существуют **чёрные**, так и мы в нашей повседневной жизни склонны быть слепы по отношению к случайностям. <...>
2. эссе (социолог, 1 курс бак)  
Во-вторых, это способ завести новые знакомства, найти друзей или даже вторую половинку **Существует** интересный факт: одна из восьми супружеских пар в США встретились именно благодаря социальным сетям. <...>
3. эссе (лингвист, 1 курс бак)  
120 **касается** функционального подхода, то кажется, что подход предлагаемый в **данной** статье, согласуется с **ним**, так как **главным** в трактовке предлога оказывается именно функциональная **нагрузка** одного аргумента на другой, а семантическим вектором предлога **подпринимается** именно **идея** функциональной доминанции объекта **бынад** некоторой областью, в которой находится или которой **тождествен X**» [Плунгян Рахилина 2000, с. <...>
4. аннотация (политолог, -)  
Исследование **представляет** информацию о речевом поведении мужчин и **женщин**, отражении пола в языковой культуре, о гендерлекте социально и культурно обусловленности общения полов). <...>
5. эссе (логист, 1 курс бак)  
Полагаю, что мотив, **подталкивающий** человека к суициду - большой запутанный клубок проблем, **из** которого он не может найти рациональный выход. <...>
6. эссе (юрист, 1 курс бак)  
Эти голоса возникают при прослушивании радио или просмотре телевизора, во время эфира **проходит** какой-то сбой, сопровождаемый помехами, сквозь

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

CoRST Result

web-corpora.net CoRST Result

Главная Поиск Новости Помощь Статистика

ru Войти

typo

Corpus total: 3677 documents, 301079 sentences, 3115212 words.

Search executed in a user-defined subcorpus of 90 documents, 4724 sentences, 83158 words.

Found: 7 documents, 8 contexts.

1

1. эссе (экономист, 1 курс бак)  
Если у человека богатый **словарный** который он приобрёл, прочитав не один десяток книг, **всегда** сможет найти синоним или иначе интерпретировать это слово. <...>
2. эссе (юрист, 1 курс бак)  
Эти голоса возникают при прослушивании радио или просмотре телевизора, во время эфира **проходит** какой-то сбой, **сопорождаемый** помехами, сквозь которые слышатся посторонние голоса. <...>
3. эссе (экономист, 1 курс бак)  
И **также** все знают, что каждый родитель будет из шкурки лезть чтобы **дитяти** закончил учебное заведение. <...>
4. эссе (экономист, 1 курс бак)  
Достаточно представить себе, **то** сделает двадцать обычных семи-восьми летних одноклассников с одним искусственным <...>
5. эссе (экономист, 1 курс бак)  
Эффективным, в силу того, что обращается к глубинному **и** желанию человека: делению всех окружающих на своих и чужих. <...>
6. эссе (юрист, 1 курс бак)  
Еще одним **критическим** фактом может **служит** следующее: в последнее время социологи стали замечать, что **абсолютно** **поменялось** **отношение** к использованию детского труда. <...>
7. эссе (экономист, 1 курс бак)  
Из-за своей разобщенности, у человека проявляются склонности к сдизму и другим асоциальным **поступка**, и можно сказать, что это и будет одним из видов





## 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.