Проект по КИЛИ

Выполнили:

Даша Деревцова (derevtsovadasha@mail.ru) и Марина Жукова(mvzhukova@edu.hse.ru)

23.10.2017

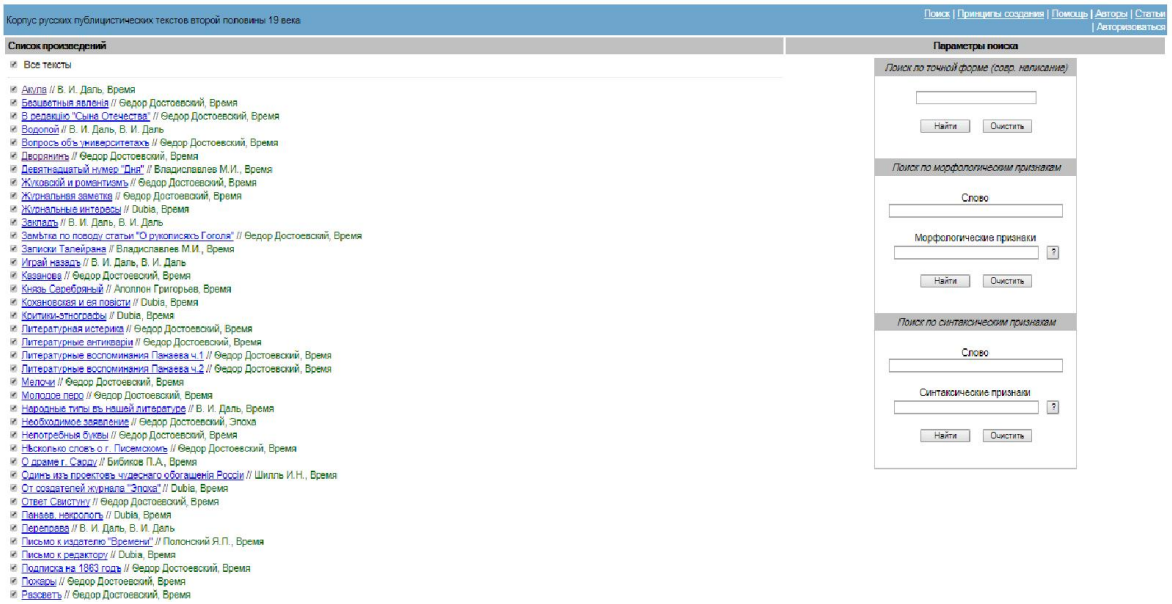
Корпус русских публицистических текстов второй половины 19 века

(http://smalt.karelia.ru/corpus/index.phtml)

Для проектной работы мы выбрали Корпус русских публицистических текстов второй половины 19 века Петрозаводского государственного университета. Так как корпус включает в себя произведения сравнительно небольшого промежутка времени, то он небольшой - всего 49 текстов. Он имеет два слоя разметки: морфологическую и синтаксическую.

Дизайн:

Оформление сайта очень спокойное и сдержанное, нет ничего лишнего, избыточного. Цветовая гамма гармоничная, не режет глаз. Главной страницей является страница поиска. Она поделена на две колонки. Слева перечислены произведения с указанием автором, к сожалению, нет даты создания. Их можно открыть и просмотреть. Большим плюсом является возможность выбрать определенные тексты для работы; можно выбрать одно, несколько, все, причем для этого есть специальное поле, где можно поставить галочку (она стоит по умолчанию). Справа находится столбец «Параметры поиска». Есть поиск по точной форме, по морфологическим признакам, по синтаксическим признакам, которые можно выбрать из предлагаемого списка.



Меню находится в правом верхнем углу, в том же блоке, что и название корпуса. Это неудобно, непривычно, но оно особо не нужно. Кнопка «Принципы создания» переводит нас в раздел, где есть ссылки на статьи по теме, но эти ссылки ведут на несуществующие страницы. Раздел «Помощь» номинально есть, но он пустой. Зато можно найти информацию о создателях корпуса, причем не только их имя-фамилию, но и род деятельности и адреса электронной почты. Также можно зайти в раздел с полезными статьями про корпусы и корпусную лингвистику, которые легко оттуда скачиваются в формате PDF. Кроме того, в меню есть кнопка «Авторизоваться», но непонятно, как это сделать, ведь нет возможности зарегистрироваться для начала.

Еще одним недостатком сайта является отсутствие привычного «подвала» внизу страницы. Из-за этого появляется ощущение незаконченности в разработке сайта.

Нельзя не отметить как достоинство адаптацию сайта при изменении масштаба или под мобильную версию. Это удобная и поэтому выигрышная деталь при взгляде на корпус со стороны.

В целом, сложно сказать, нравится ли «внешний вид» и устройство корпуса. С одной стороны он покоряет лаконичностью и простотой, с другой - нерабочие ссылки и неудобная разметка сайта сильно разочаровывают.

Onboarding:

Данный ресурс найти легко - это первая ссылка при поиске по названию корпуса. Более того можно перейти в этот корпус через ссылку в НКРЯ. Форму поиска искать не надо, так как, мы уже отмечали, она открывается как главная страница. Все поля поиска подписаны, поэтому ориентироваться, как и по каким параметрам искать, крайне просто. Можно с уверенностью сказать, что сайт интуитивно очевиден и понятен, даже несмотря на отсутствие примеров и образцов запросов.

Помощь пользователю:

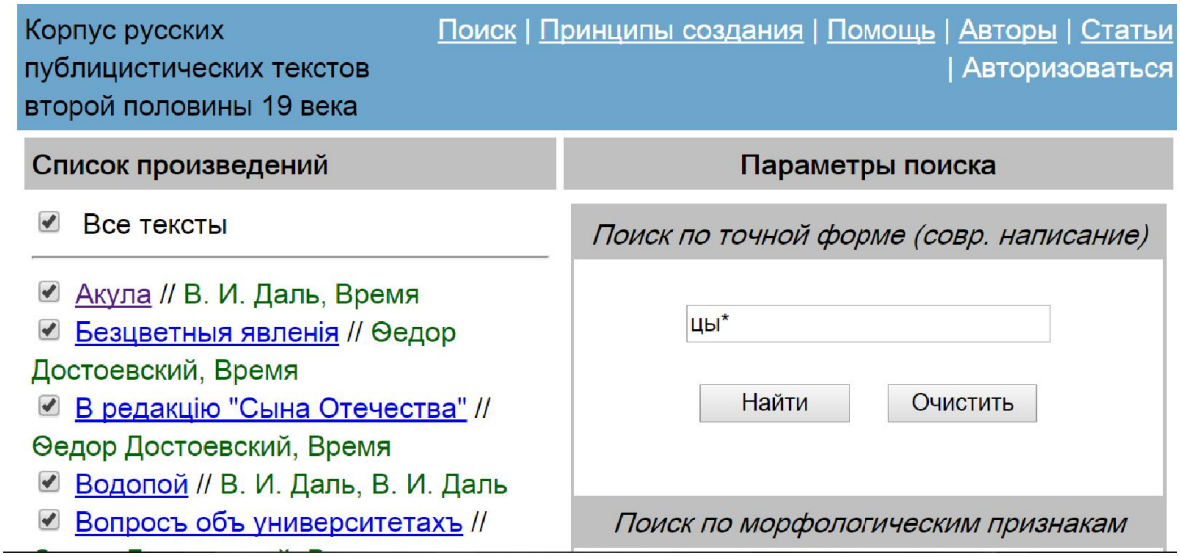
Как уже было сказано выше, на сайте отсутствуют любые подсказки, life hacks и инструкции для начинающих пользователей. Несомненно, по этому критерию данный корпус проигрывает многим другим, но так как он понятен интуитивно, это небольшая проблема. Более серьезным минусом является отсутствие описания корпуса и его структуры, поэтому пользователю сложно получить целостное представление о данном ресурсе.

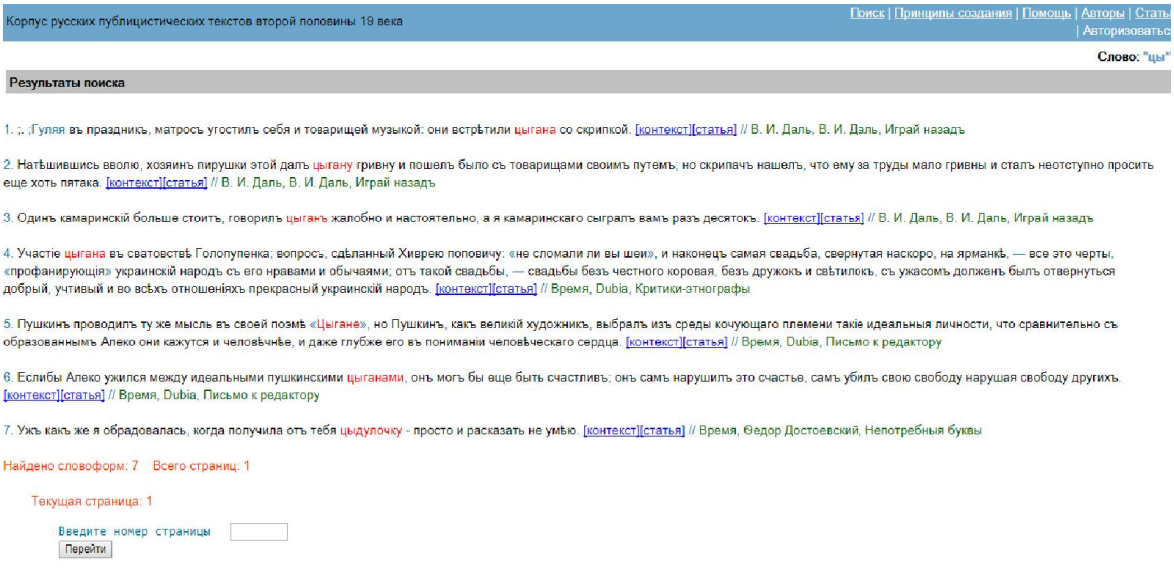
Примеры запросов:

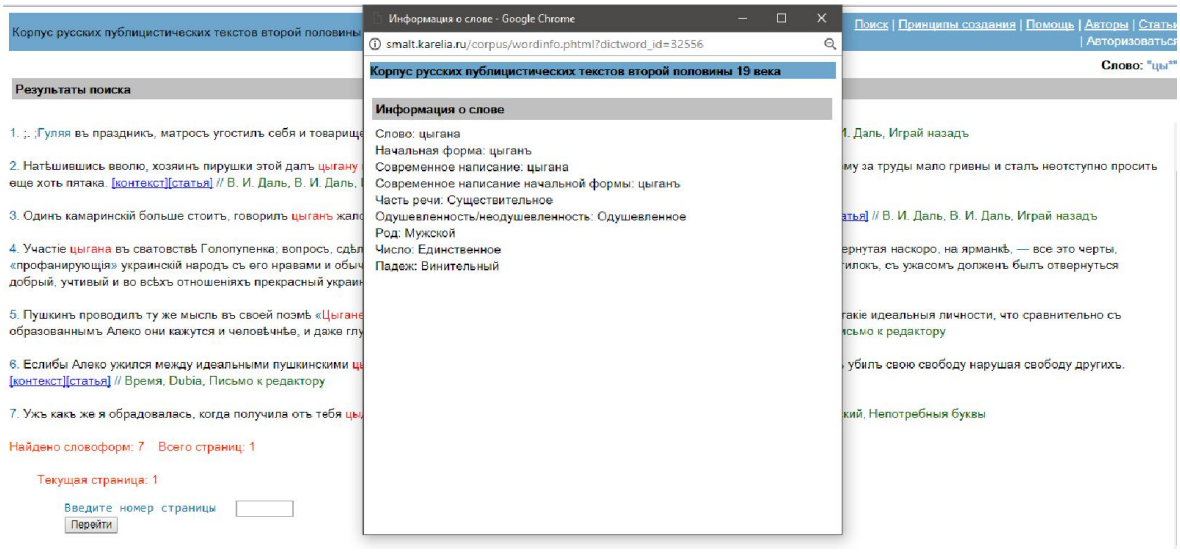
Теперь мы переходим непосредственно к практическому использованию данного корпуса для лингвистических исследований, его функционалу и

возможностям. Надо сразу отметить, что в текстах используется старая орфография; в поисковое поле же надо вводить слово (это указано, что является достоинством) с современным написанием(корпус учитывает данный факт).

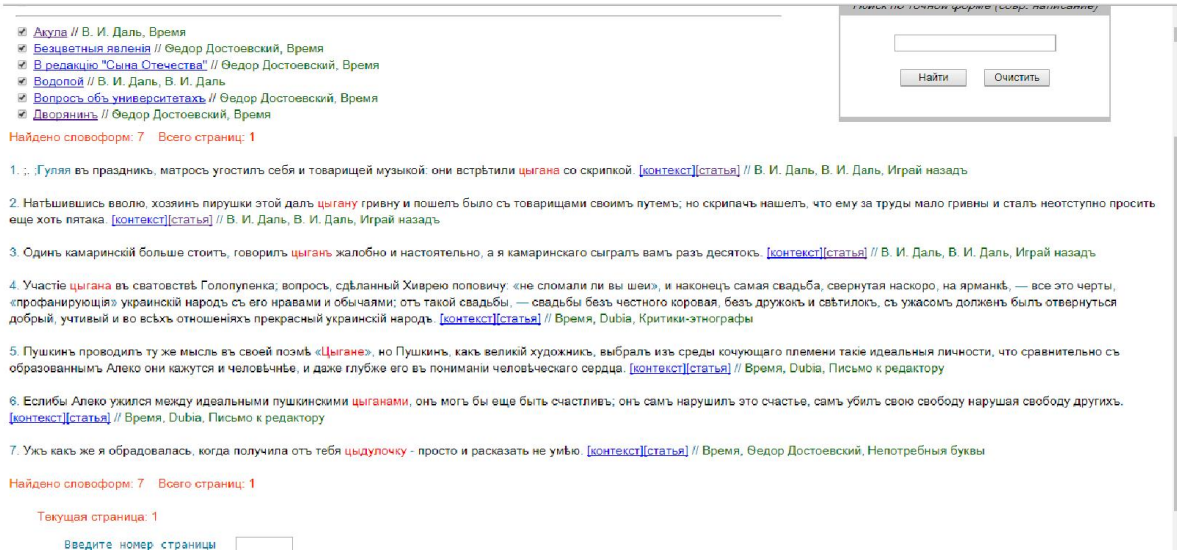
1) Для начала проверим корпус на поиск слов по маске. Введем маску «цы*» в поле для поиска по точной форме. Результат выглядит как список текстов, где встречается слово, начинающееся с «цы-». Указаны авторы и названия текстов. Нужное слово выделено цветом, при нажатии на него открывается окно с информацией о слове, то есть его категории, в том числе указано его устаревшее написание. Также можно просмотреть контекст искомого слова или же прочитать произведение целиком, что очень удобно.



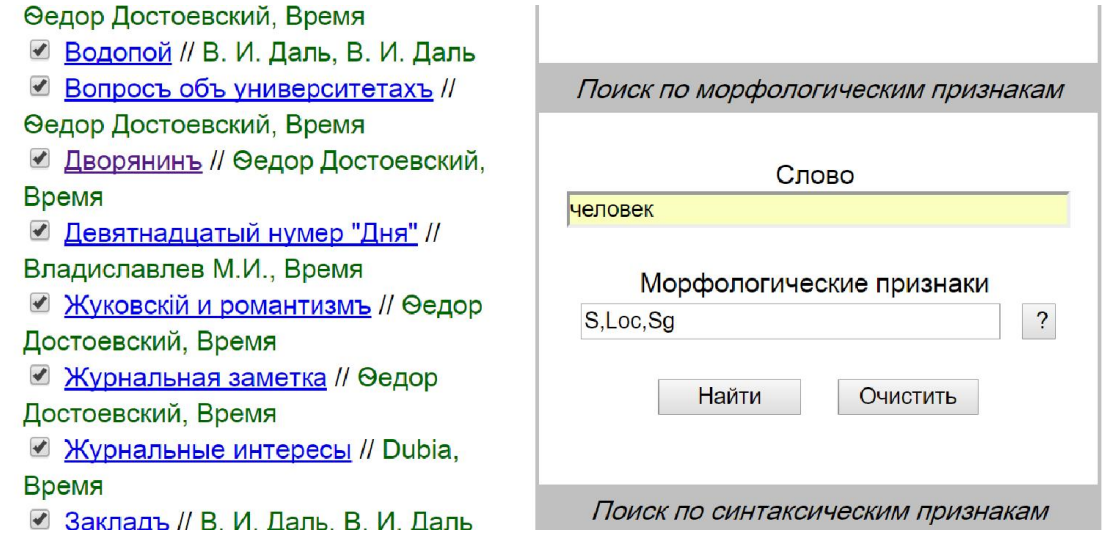


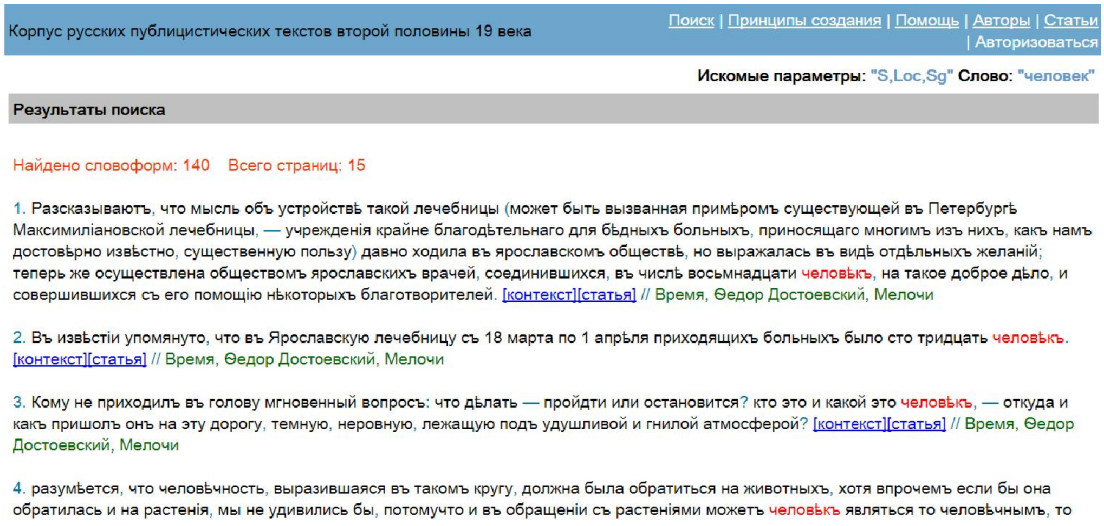


При нажатии на кнопку «Поиск» с намерением возвратиться на главную страницу видишь “интересную” картину: ты остаешься на той же странице с результатами выдачи, но поиск открывается на половину окна, то есть экран становится как бы поделенным горизонтальной линией- сверху поиск, снизу результат.

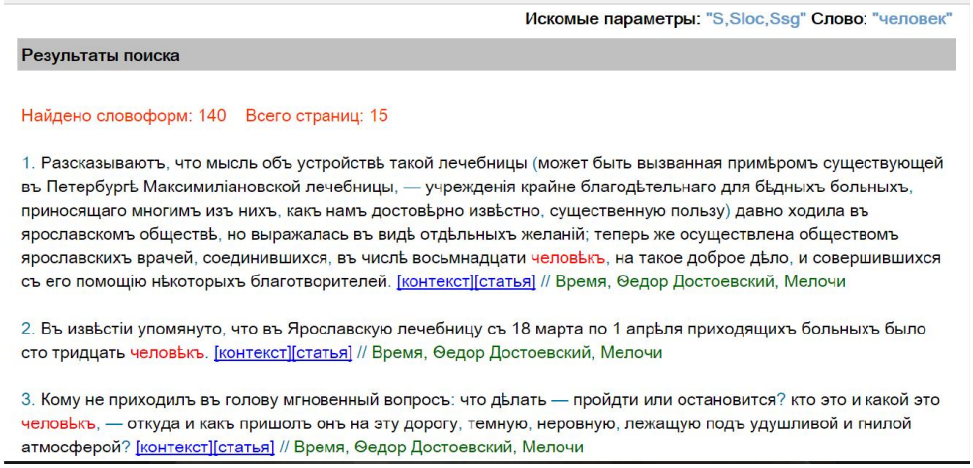


2) Теперь введем запрос сложнее. Посмотрим, сколько раз в корпусе встречается слово «человек» в форме предложного падежа единственного числа. Получаем, что данная словоформа встречается в корпусе 140 раз.



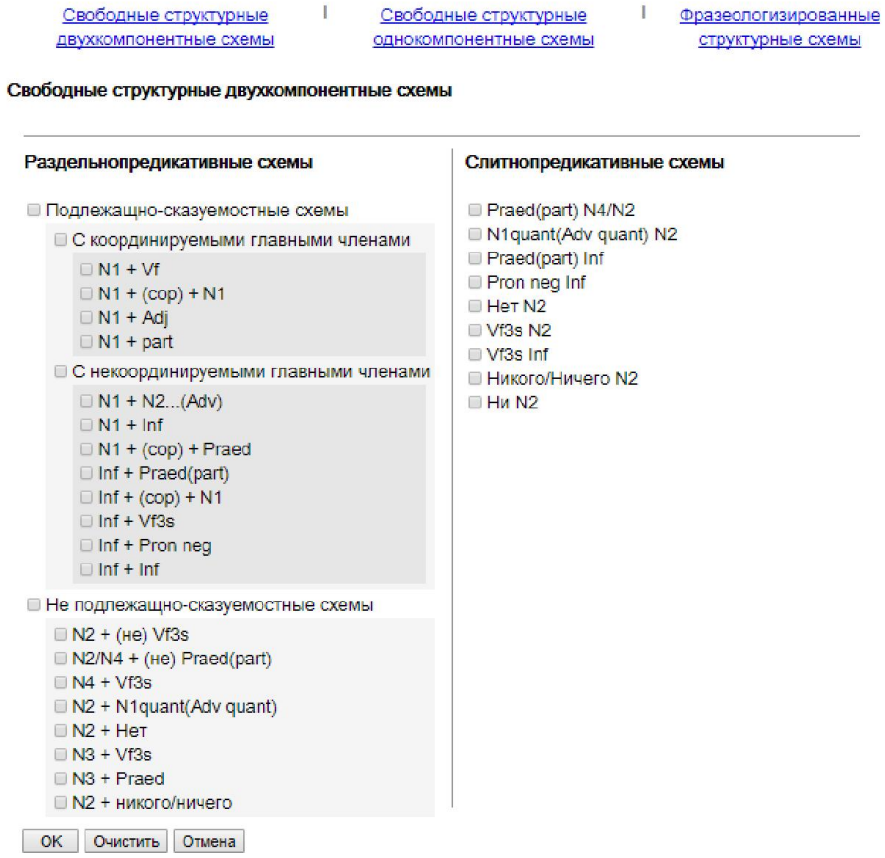


Но если просматривать результаты, то мы видим, что выданные словоформы не соответствуют запросу. Самое интересное, что морфологическая разметка неподходящих нам слов в выдаче полностью правильная: например, в первом тексте словоформа «человек» размечена как форма генитива множественного числа. Потом мы меняем морфологические признаки того же самого слова, но результат получается таким же.

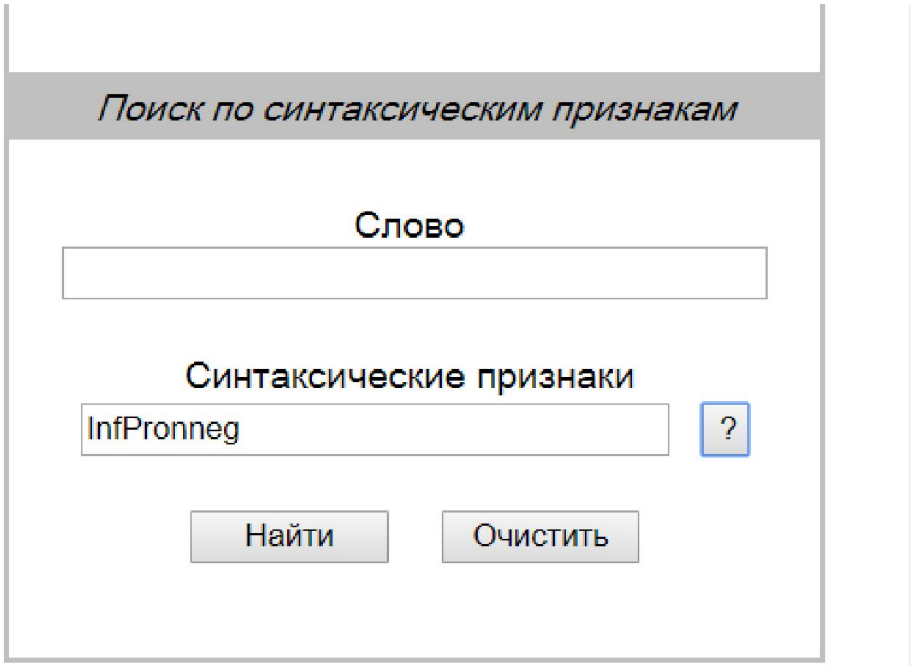


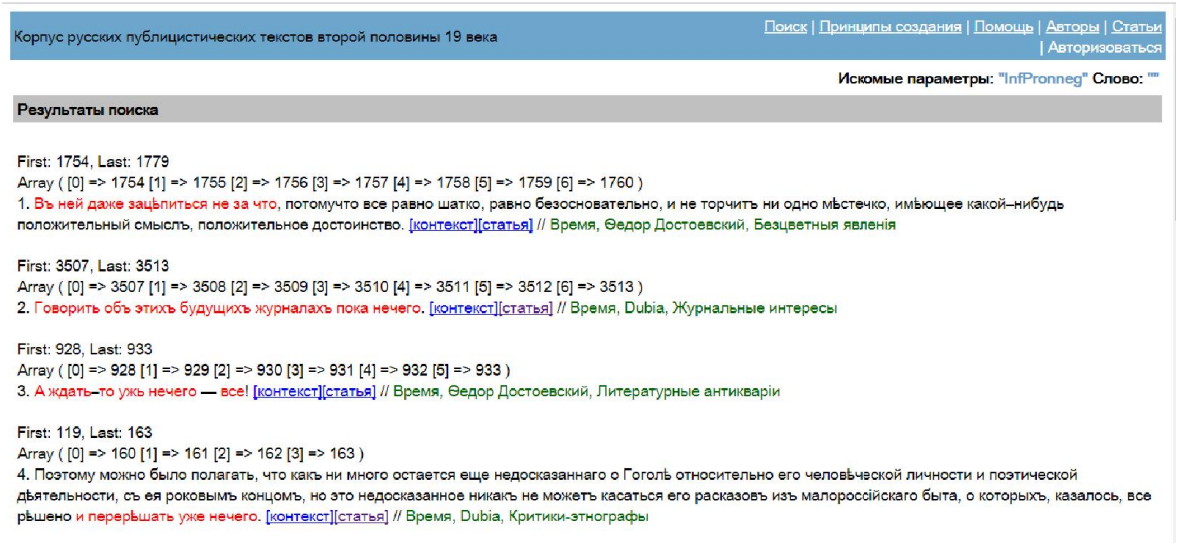
Из этого мы делаем вывод (проверив догадку еще на паре других слов), что поиск по морфологическим параметрам не работает, несмотря на морфологически правильную разметку слов.

3) Наконец, проверим, как работает поиск по синтаксическим признакам. Надо сказать, что под синтаксическими признаками подразумевается вхождение слова в некоторые конструкции, а не его роль в предложении. Хочется сразу заметить, что свод данных параметров кажется удобным, так как все признаки разделены на три категории: « Свободные структурные двухкомпонентные схемы», «Свободные структурные однокомпонентные схемы», « Фразеологизированные структурные схемы»,- которые предлагают свои варианты синтаксических конструкций. При наведении на определенную синтаксическую схему появляется пример предложения с данной конструкцией, это very useful.



Посмотрим, сколько результатов будет при выборе раздельнопредикативной схемы с некоординируемыми главными членами «Инфинитив + отрицательное местоимение» без ввода определенного слова.





Всего 14 примеров, причем все они удовлетворяют параметрам поиска. Но сразу видно, что помимо самих примеров есть лишние строки «First» и «Array», которые являются доказательством недоработки сайта.

К сожалению, не на все синтаксические схемы есть примеры из корпуса. Так, там нет ни одной конструкции, построенной по фразеологизированной схеме.

Достоинства и недостатки:

Все достоинства и недостатки данного корпуса можно представить в виде таблицы:

Benefits	**Drawbacks**
Приятный дизайн	Сайт иногда не работает, «вылетает»

Страница поиска открывается по умолчанию	Незаконченная разметка сайта и неудобное расположение меню
Есть возможность выбрать произведения для работы	Много нерабочих ссылок и пустых разделов
Работает поиск по маске слова	Нет инструкций, примеров запросов
Полные списки морфологических и синтаксических признаков	Не работает поиск по морфологическим признакам
	Нет примеров фразеологизированных конструкций
	Невозможно искать сразу по морфологическим и синтаксическим параметрам
	В списке морфологических параметров можно выбрать несочетаемые признаки (например, возвратность-невозвратность глагола)
	Нет статистики употребления того или иного слова
	Нет метаразметки, нет возможности задавать подкорпус

Вывод:

В целом, корпус кажется незаконченным с точки зрения как разработки сайта, так и его состава. Кажущиеся на первый взгляд удобные методы поиска оказываются неэффективными, вдобавок сам поиск не всегда работает. Довольно внушительное количество недостатков корпуса и маленький объем не позволяют использовать его для лингвистических исследований.