**Беларускі N-корпус**

[**http://bnkorpus.info/**](http://bnkorpus.info/)

**Основная информация**

Белорусский N-корпус – первая попытка создания публичного общего корпуса белорусского языка. Содержит в обработанном виде 185 097 текстов, 79 382 061 слов, в необработанном – 22 244 текстов и 208 474 017 слов.

Есть доступ онлайн, также доступны для скачивания программы проверки орфографии для Microsoft Office/LibreOffice/OpenOffice/Mozilla Firefox/Thunderbird. Есть zip-архив с лексико-грамматической базой.

Последняя редакция корпуса проводилась 14.08.2013.

**Дизайн**

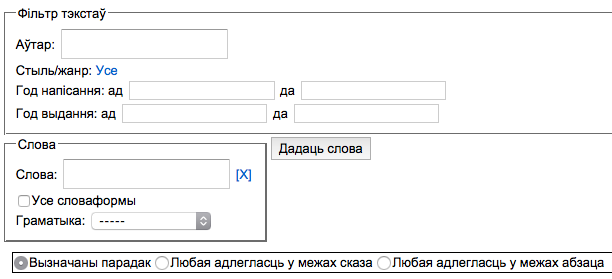
Этим атрибутом белорусский корпус похвастаться не может, потому что дизайна там просто нет! Светлый белый фон, черные буковки на нем…

Ориентироваться в корпусе можно без проблем, так как выбранный разработчиками минималистический дизайн не отвлекает от главной задачи поиска.

Все те же разработчики приняли странное решение расположить весь текст с левой стороны страницы, при этом середина абсолютно пустая и ничем не занята (а хорошо бы!). Также мешают взгляду и отвлекают на себя горизонтальные линии через весь экран, не несущие никакой смысловой нагрузки.



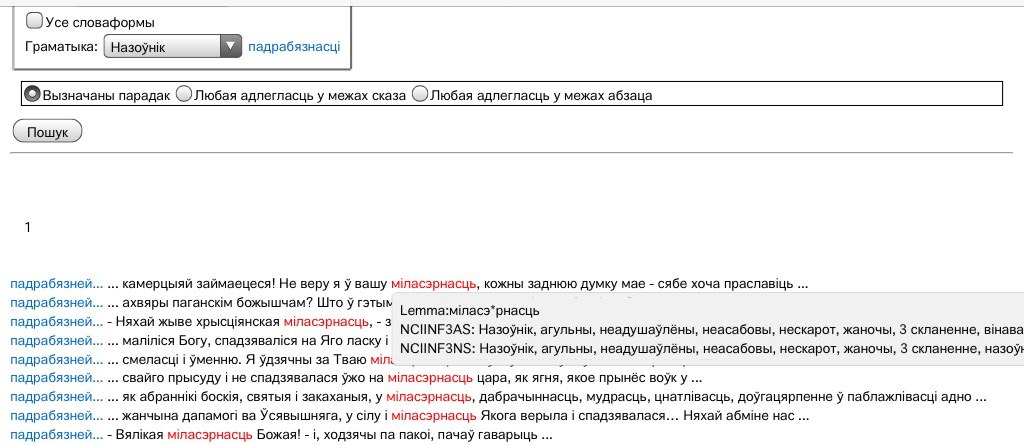
Текст однороден, все расположено рядом и плохо различимо. С таким количеством свободного пространства запихивать весь текст в один угол и так плотно – не самая рациональная идея.



**Разметка корпуса**

Корпус размечен по следующим параметрам: лемматизация (указание для каждой словоформы из текста ее исходной формы), морфо-синтаксическая разметка (выделение основ, определение части речи и признаков грамматических категорий).

Морфологической разметки в корпусе нет. Аффиксы не выделены, при поиске сложных слов никак не указано, что они сложные.

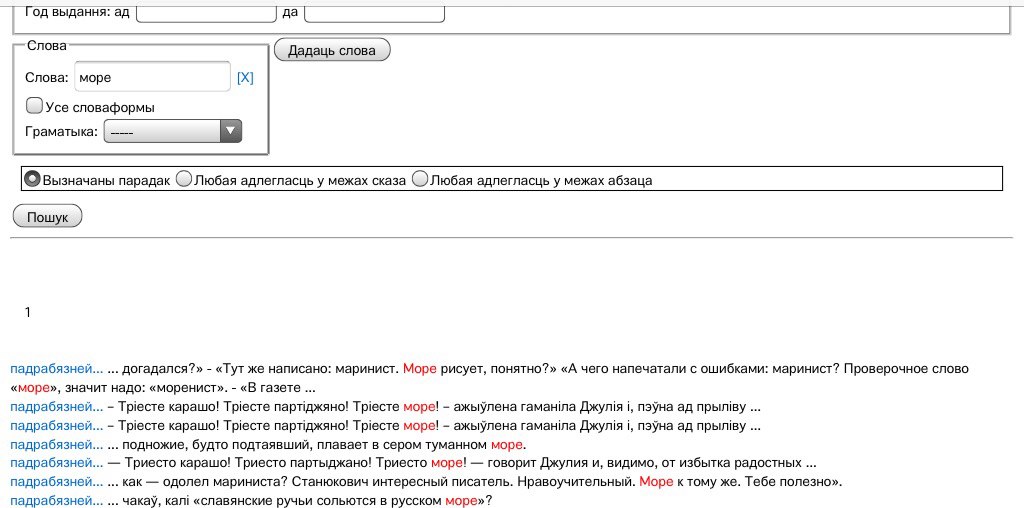


Синтаксическая разметка отсутствует. Невозможно узнать ни тип предложения, ни то, каким членом предложения является какое-нибудь слово. Характер синтаксической связи также не указан.

Семантическая омонимия не снята. При нажатии кнопки «подробнее» появляется сразу несколько характеристик слова, которое искал пользователь. Эта информация к тому же загораживает само предложение, членом которого является рассматриваемое слово.



Дискурсивная разметка есть, но не имеет общего вида. В каждом тексте используются разные знаки препинания при репликах.

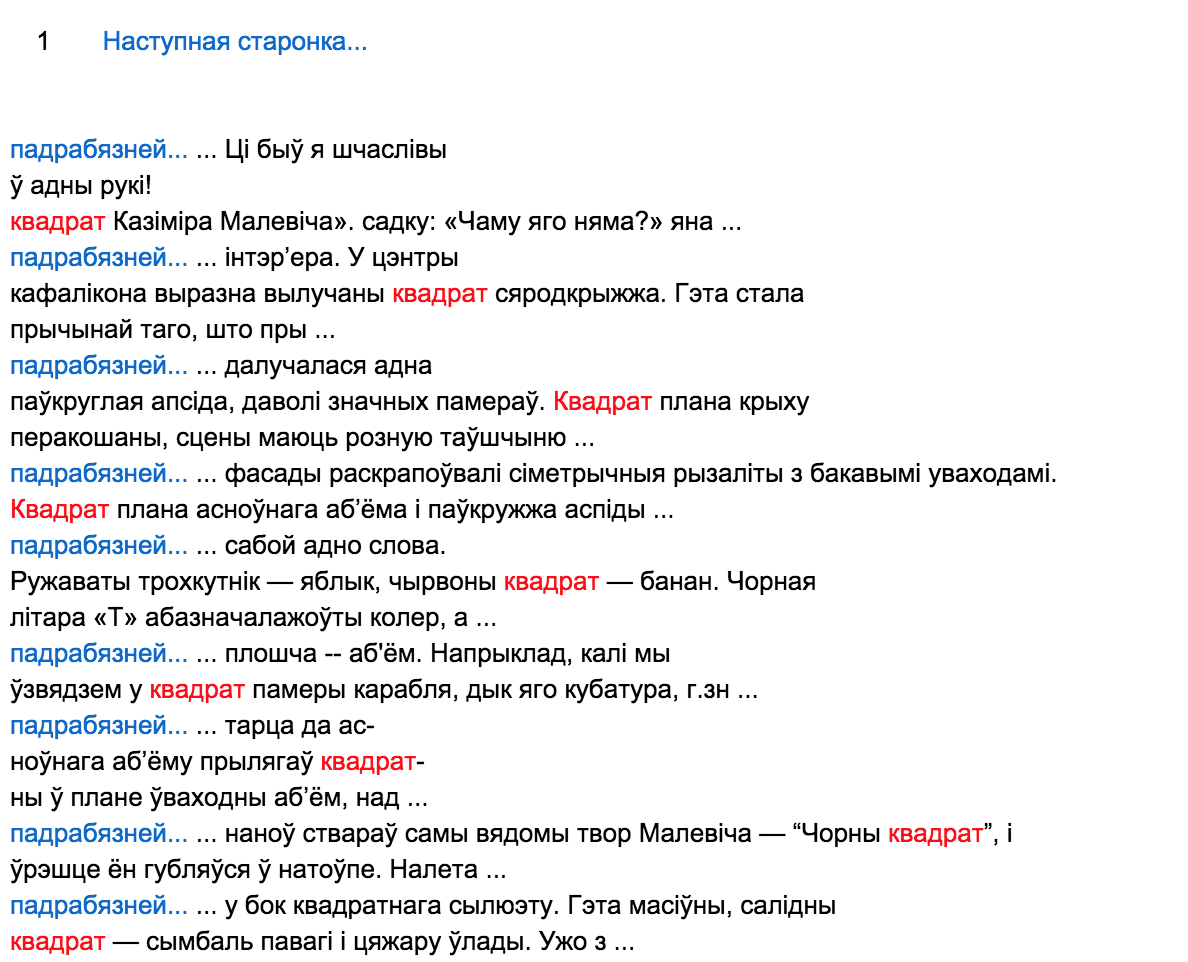
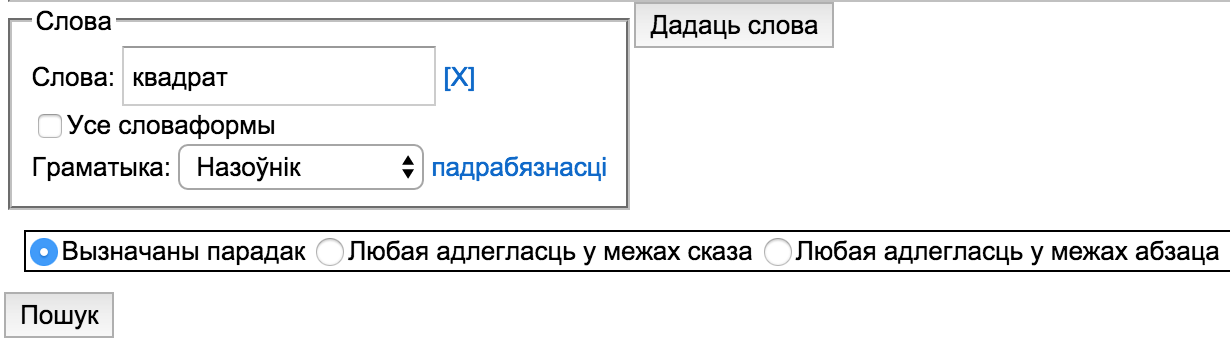


Существует подкорпус необработанных текстов, который представляет собой сбор текстов, которые были распознаны в автоматическом режиме и не прошли вычитку (есть проблемы с переносами, разделением на предложения, абзацы и пр.).Подкорпус позволяет искать слова и словоформы, но не имеет полноценной метаразметки.

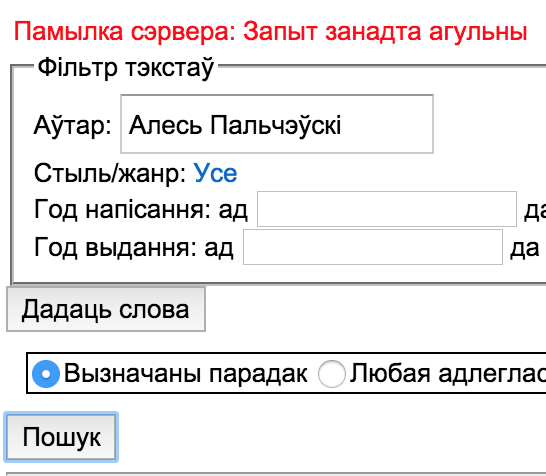
**Инструкции и туториалы по корпусу**

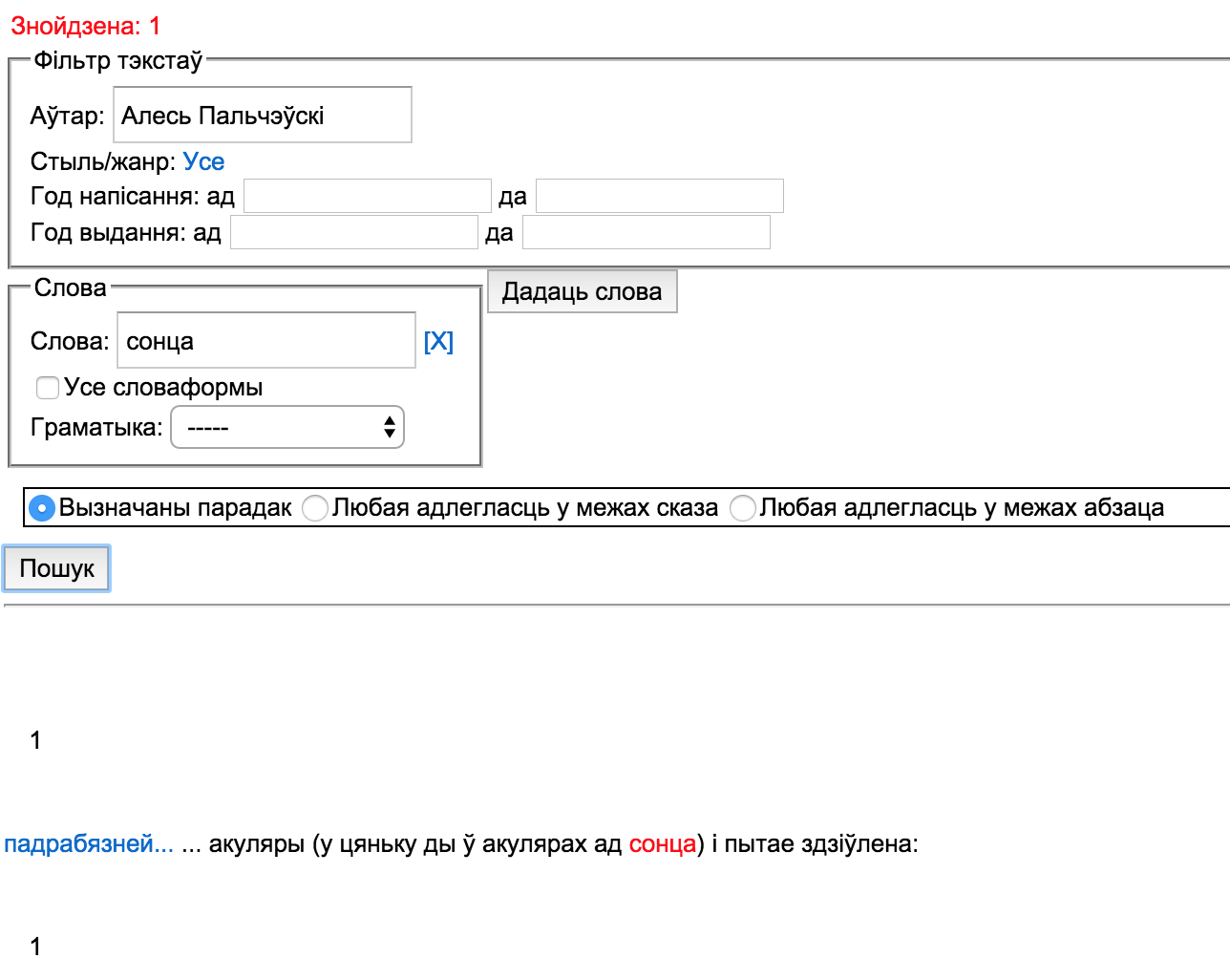
Инструкций на сайте не наблюдается. Информации о корпусе в Интернете нет, видео-туториалов нет.

**Функционал, примеры запросов**

**1.** Проведем запрос текстов, в которых встречается существительное "квадрат". Вводим в поле поиска слово, в "грамматике" выбираем существительное. Видим, что найдено более 100 результатов, однако сколько именно – неизвестно. Также при показе результатов не выдается общее количество результатов, а также количество страниц. До тех пор, пока мы не перейдем на вторую страницу, мы не узнаем, существует ли третья. Это явно неудобно для пользователя, особенно для того, который пользуется корпусом впервые.

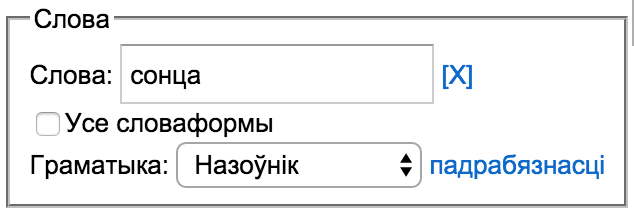
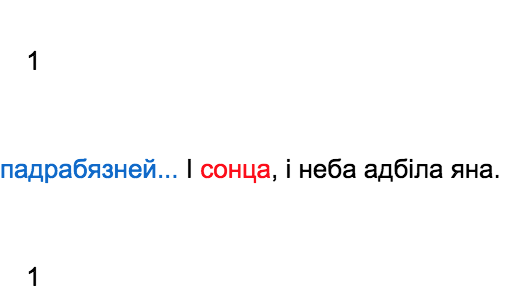
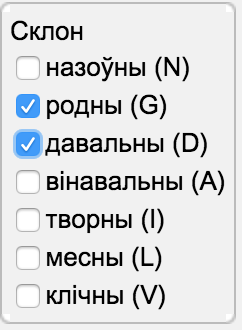
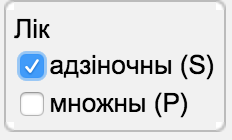
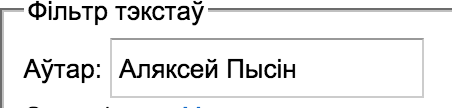


**2.** Попробуем найти тексты определенного автора с указанием какого-либо слова. Просто при вводе любого автора корпус выдает: "Ошибка сервера: Запрос слишком общий".

Остановимся на авторе Алесь Пальчэўскі и упоминании в его текстах слова "сонца" (солнце). Получаем 1 результат.

**3.** Сделаем запрос более конкретным: слово "сонца" в текстах автора Аляксей Пысiн, употребленное в родительном или дательном падежах в единственном числе.

Получаем единственный результат.



При попытке уточнить годы написания произведения, корпус не выдает результатов, видимо, потому, что не указан сам год написания произведения. Хотя в случае с другими авторами часто выдает "ошибку сервера".

**Вывод:** неполная разметка корпуса и, видимо, не самые стабильные сервера дают возможность делать несложные поисковые запросы, но не позволяют искать большие объемы данных (например, все тексты автора) и искать, используя большое число критериев.