Национальный корпус польского языка
Narodowy Korpus Języka Polskiego

Маслова Мария (mashamaslova85@gmail.com)

Гребёнкина Мария (ya.grebenkina-maria-74@ya.ru)

Кошкарёва Диана (di.coshckarewa@yandex.ru)

Миронов Михаил (orderlyoftime@gmail.com)

1. Данные о ресурсе

Национальный корпус польского языка (Narodowy Korpus Języka Polskiego) – это интернет-ресурс, основанный на собрании текстов на польском языке. Был создан в 2008 году. Делится на два механизма поиска: Поликарп и Пелкра. Первый работает на двух языках: польский и английский, однако, к сожалению, второй работает только на польском.

Адрес: http://www.nkjp.pl/

2. Дизайн и устройство корпуса

Шрифт, преимущественно используемый в корпусе, – Times New Roman, встречающийся повсеместно благодаря тому, что считается в некотором роде классическим.

Цветовая гамма сайта достаточно приятна на вид. Сочетание серого, белого и красного не раздражает взгляд, а белый и красный к тому же отсылают к цветам польского флага, что демонстрирует внимание разработчиков к деталям.

Интерфейс корпуса вполне удобен. Слева сгруппированы ссылки на материалы об истории корпуса, команде разработчиков, публикациях, благодарностях и тому подобных вещах, а также ссылка непосредственно для работы с корпусом, причем доступна она только в англоязычной версии – в польской по этому же адресу находятся сведения об использовании корпуса в различных проектах, что несколько сбивает с толку. Впрочем, вне зависимости от выбранного языка, ссылки для работы будут находиться справа.

Нарекания вызывает в первую очередь страница с подсказками, у которой отсутствует какое бы то ни было оформление – это просто текст. Впрочем, найти их достаточно легко, а это все-таки важнее. Ссылка на страницу находится слева от меню поиска, подсказки касаются в основном разметки, используемой в корпусе, и того, как следует оформлять запросы.

Версия сайта для мобильных устройств работает корректно, но вот десктопная подкачала: к сожалению, в ней не всегда работает вывод диакритических знаков.

3. Глазами новичка

***Взгляд 1**: Корпус довольно легко найти: достаточно выполнить запрос в любой поисковой системе. Можно заметить, что на сайте НКРЯ есть ссылка на корпус*

польского. Это делает поиск данного корпуса гораздо легче для русскоговорящих пользователей. Для англоговорящих пользователей поиск тоже достаточно прост. Сайт функционирует на двух языках: английском и польском. Форму поиска можно найти на всех страницах сайта, она находится в меню справа. На самой странице поиска под полем ввода есть специальные клавиши с польскими буквами, которые могут не присутствовать на клавиатуре пользователя. Слева находится небольшое меню с пунктами: поиск, настройки, сообщить об ошибке, помощь. В разделе "помощь" находится достаточно подробное описание работы и элементарные примеры запросов. Поиск осуществляется достаточно быстро, особенно для простых запросов.

***Взгляд 2:** Корпус симпатично выглядит и красиво оформлен. Возможен ввод диакритических символов при отсутствии таковых на клавиатуре. Можно работать и с мобильного телефона. Сам процесс несложен, однако работу с пелкрой сильно затрудняет то, что она только на польском. Результаты ищутся довольно быстро, но смена настроек иногда проходит затруднительно. Интересно наличие «слов дня» - самых частотных слов за последний день или неделю.*

4. Продвинутый функционал (поиск по корпусу)

Ресурс позволяет осуществлять поиск словоформ, лексем, поиск по грамматическим категориям и метатекстовым данным. Семантическая, синтаксическая разметка отсутствует.

В строку ввода можно вбивать непосредственно словоформу (дополняя запрос спец. символами) или осуществлять поиск при помощи атрибутов:

- base задаёт поиск словоформ по лексеме
- orth задаёт поиск определённой словоформы или *сегмента*
- pos задаёт поиск словоформ по грамматическому значению
- Возможно использовать как атрибут название грамматической категории

attribute	possible values
number	sg pl
case	nom gen dat acc inst loc voc
gender	m1 m2 m3 f n
person	pri sec ter
degree	pos comp sup
aspect	imperf perf
negation	aff neg
accentability	akc nakc
post-prepositionality	npraep praep
accommodability	congr rec
agglutination	agl nagl
vocalicity	nwok wok
fullstoppedness	pun npun

Для создания сложных запросов можно использовать специальные символы и цифры:
`?`, `*`, `+`, `.`, `,`, `|`, `{`, `}`, `[`, `]`, `(`, `)`

- | - логическое ИЛИ. Результатом запроса "Ala|Ela" будет *Ala* или *Ela*,
- Внутри квадратных скобок заключены эквивалентные друг другу символы, то есть результат запроса [AE]la будет таким же, как и в предыдущем пункте
- Знак вопроса маркирует предшествующий символ как необязательный, то есть результатом запроса "beza?" будет и *bez, и beza*,
- Точка заменяет один символ (но не его отсутствие). Выдача по запросу "bez." включит *beza*, *bezy*, *bezą*, и т. д., но не *bez* или *bezami*,
- Астерикс означает, что в выдачу будут включены слова с любым количеством символов, предшествующих астериксу, в указанном месте. Например, по запросу "a*by" будут выданы варианты *by* (ноль *a*), *aby*, *aaaaby*, и т. д.,
- Сочетание .* заменяет любое количество любых символов. В выдачу по запросу "Ala.*" будут включены все слова, начинающиеся с *Ala*: *Ala, Alabama и т. п.*,
- Плюс по значению равен астериксу, но число заменяемых символов строго больше ноля.

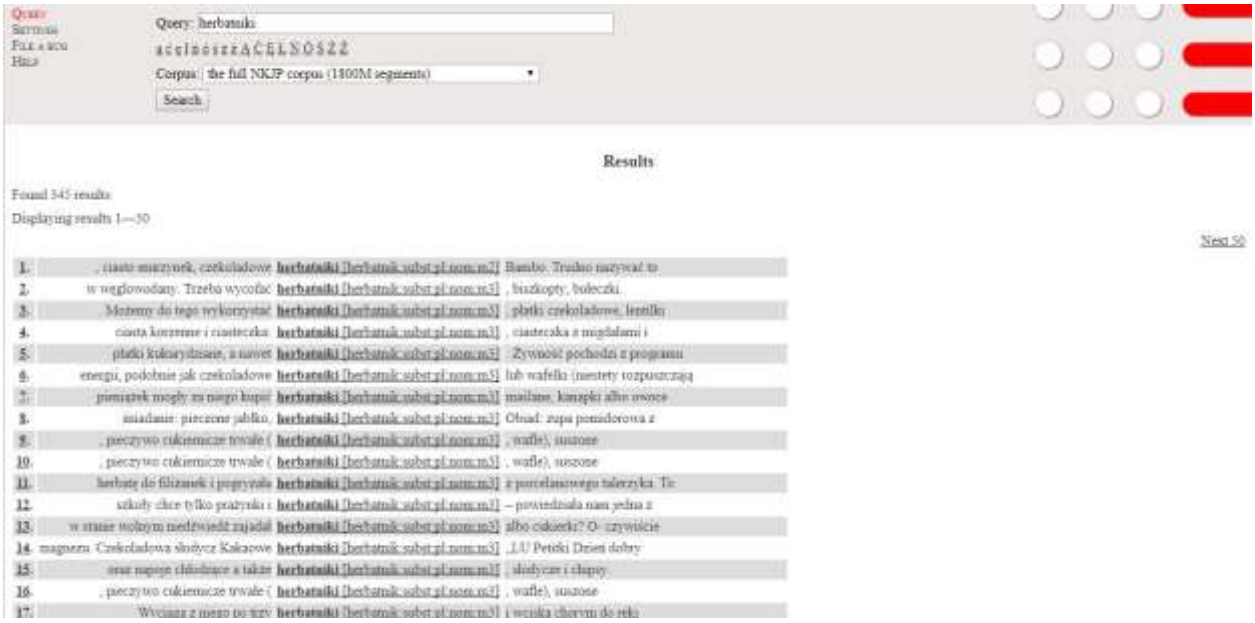
- Выражение {n,m} означает от n до m повторений предшествующего символа. Результатом запроса "a{1,3}b.*" будут словоформы, содержащие от одной до трех a перед b: *aby*, *aaaby*, *absolutnie*, и т. д. Результатом запроса ".*(la){3,}.*" будут выражения, содержащие как минимум три *la* подряд: *tralalala*, *sialalala*,
- Сочетание /i в конце запроса определяет нечувствительность к регистру
- & - логическое И
- ! – отрицание

Poliqarp также предоставляет возможность поиска определённой фразы только *внутри одного* предложения или абзаца. Для этого используются определители within s и within p соответственно.

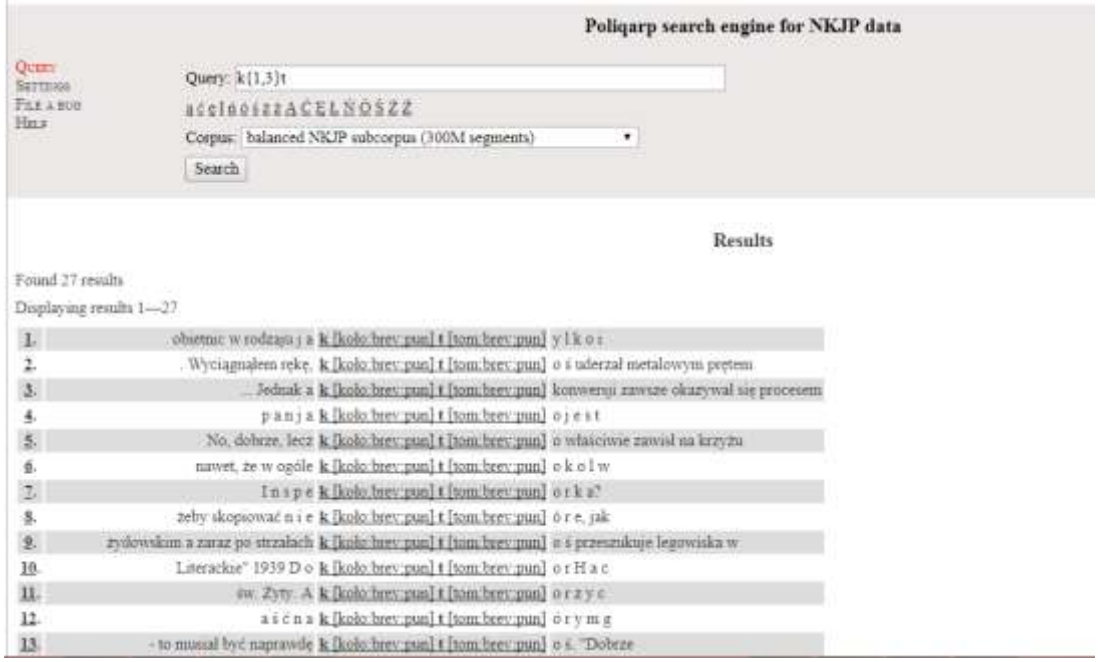
Помимо прочего, с помощью системы атрибутов можно задавать значения метатекстовых атрибутов, таких как имя автора, год создания, источник и др. Например, запрос “[pos=verb] meta channel=mowiony” выдаст все глаголы из текстов-записей устной речи.

Преимуществом поисковой системы является возможность дальнейшей офлайн работы с выдачей корпуса: результаты легко загружаются в excel или html.

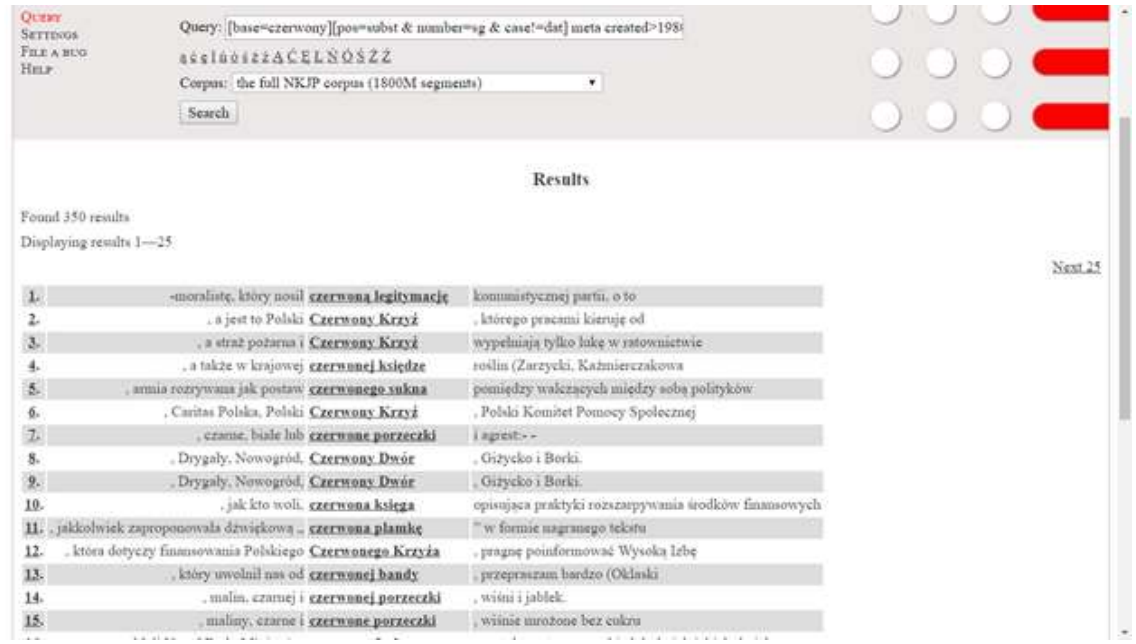
Пример запроса: *Слово herbatniki (печенье)*



Пример запроса: *От одной до трёх букв k перед t*



Пример запроса:
Все словоформы czerwony (красный), за которыми следует существительное единственного числа не в дательном падеже, из текстов, созданных позднее 1980 года
[base=czerwony][pos=subst & number=sg & case!=dat] meta created>1980



Плюсы:

- Дизайн корпуса гармоничен и приятен на вид
- Имеется возможность скачивания выдачи
- Интересные «фишки»: слова дня, побуквенный поиск
- Возможность работать на двух языках
- Панель диакритики

Минусы:

- Путаница в использовании английской и польской версии; слабо проработанная английская
- Медленная скорость загрузки, особенно при изменении настроек