Описание лингвистического компьютерного ресурса

Алексей Старченко (aleksey-starchenko@mail.ru),
Алина Тиллабаева (alinka99-t@mail.ru),
Элина Новопашенная(elinanovo@gmail.com)

Содержание

1 Данные о ресурсе

Выбранный ресурс - Польско-русский и русско-польский параллельный корпус (http://pol-ros.polon.uw.edu.pl). Ресурс предоставляет возможность онлайн-поиска по корпусу размером около 30 млн. слов.[1] В корпусе представлены художественные (русские, польские и переводные с других языков), юридические, религиозные тексты и пресса.

2 Дизайн

При первом взгляде корпус создаёт хорошее впечатление. Цветовая гамма, за исключением бордовой линии внизу страницы с данными о создателях сайта и финансировании проекта, светлая, приятная для глаза. Навигация по сайту удобная: в важные разделы можно попасть из бокового меню.

[1] Подробнее о составе корпуса см. Приложение.

Все кнопки достаточно большого размера, их расположение логично организовано. Есть возможность набора текста с виртуальной клавиатуры. Достаточно неудобным для использования является список грамматических категорий, представляющий из себя очень длинную выдвижную панель. Хотя это и объясняется большим количеством категорий, которые нужно выводить для двух языков, можно было бы скомпоновать их более компактно.

Сильным недостатком сайта является очень медленная скорость выполнения запросов.

3 Ресурс глазами новичка

Найти ресурс можно без труда с помощью поисковых систем, а форму поиска - с помощью бокового меню.

Сайт доступен на русском, польском и английском языках. Названия разделов, кнопок, окон ввода и грамматических категорий подписаны (подписи выполнены качественно по крайней мере на английском и русском языках).

Интерфейс поиска интуитивно понятен. Задание параметров (грамматических категорий, языка произведения и т.д.) происходит с помощью выбора мышью, а не с помощью тэгов, что гораздо проще, хотя и усложняет дизайн. Единственный неочевидный момент - где находится поиск по точной формуле, а где по лемме, однако это можно узнать в разделе «Как работает корпус». В этом же разделе можно найти примеры запросов.

4 Помощь пользователю

Для помощи пользователю на сайте корпуса есть страница «Как работает корпус». Все подсказки и инструкции собраны в одном месте, что весьма удобно. Раздел достаточно небольшой, но содержит почти всю необходимую информацию, так что с ней можно быстро ознакомиться.

На других ресурсах обзоров по использованию корпуса нет, но это компенсируется интуитивно понятным интерфейсом, схожим с интерфейсом НКРЯ и наличием инструкций на сайте корпуса.

5 Продвинутый функционал

Параллельные корпуса представляет большой интерес для научно-исследовательской работы: для поиска различий в грамматической системе языков, для исследований языковых картин мира через контексты употребления слов, для обучения систем машинного перевода и так далее. Польско-русский параллельный корпус особенно интересен, так

как можно наблюдать, какие различия появились между не так давно, но уже заметно разошедшимися родственными языками. К сожалению, данный ресурс не даёт возможности выгрузить данные для дальнейшей обработки, однако возможности онлайн-поиска достаточно высоки.

Корпус даёт возможность выбрать один из предложенных подкорпусов, задать язык оригинала и время создания художественного текста, однако последнее - с достаточно небольшой точностью.

Поиск можно вести и на русском, и на польском по точной форме, по части слова, по лемме и по морфологическим признакам. Также есть возможность параллельного поиска: выдаются примеры, у которых определённые слова разных языков содержатся (или одно содержится, другое не содержится) в оригинале и переводе. Увы, отсутствует возможность морфологического параллельного поиска.

Данные инструменты позволяют проводить достаточно интересные исследования (см. пример небольшого исследования соответствия русских слов «изба» и «хата» и их польских коррелятов в разделе 6). Стоит однако заметить, что функционал корпуса заметно меньше, чем у ведущих корпусов, например, у НКРЯ: отсутствует разметка по семантическим признакам, расположению относительно знаков препинания, синтаксическая разметка; нет возможности искать по словообразующим аффиксам. Нет возможности сортировать результаты поиска, что создаёт достаточно большое неудобство, с учётом того что выдачу нельзя скачать. Кроме того, при работе корпуса иногда возникают ошибки: некоторые тексты выдаются несколько раз на один запрос, не полностью работает поиск по части слова на русском языке, некоторые слова, которые очевидно должны быть в корпусе, им не находятся, например, «я», «być».

Большим преимуществом корпуса является снятая омонимия в большей части польских текстов, однако в русских, увы, дизамбигуация не произведена.

6 Примеры поиска

Для начала попробуем поискать по морфологическим признакам. Выберем для поиска непереходные глаголы совершенного вида в мужском роде. Результат:

Poszukiwane wyrażenie | Поиск для:

Wyników | Результатов: 106329 || **Na stronę | На странице:** 10 || **Tryb:** wyszukiwanie wśród publikacji w dostępie ogólnym | **Режим**: Поиск публикации в общем доступе

1	Тренироваться начали сразу же, как он у нас появился. [Догги-стайл на ринге, Жеглая Ксения]	Treningi zaczęliśmy od razu, gdy zjawił się u nas w domu. [Tańczący z psami, Żeglaja Ksenia]
2	— Недавно во время выступления Галины Чоговадзе из зрительского ряда выбежал лабрадор и пристроился к танцующим. [Догги-стайл на ринге, Жеглая Ксения]	Niedawno w czasie występu zawodniczki Galiny Czogowadze z widowni wpadł na parkiet wielki labrador i zaczął tańczyć razem z para zawodników. [Tańczący z psami, Żeglaja Ksenia]
3	— Недавно во время выступления Галины Чоговадзе из зрительского ряда выбежал лабрадор и пристроился к танцующим. [Догги-стайл на ринге, Жеглая Ксения]	Niedawno w czasie występu zawodniczki Galiny Czogowadze z widowni wpadł na parkiet wielki labrador i zaczął tańczyć razem z para zawodników. [Tańczący z psami, Żeglaja Ksenia]
4	Подпоручик Янек Яжембский в условленное место прибыл пунктуально. [Стечение обстоятельств, Хмелевская И.]	Podporucznik Tadzio Jarzębski, piastujący stanowisko podkomisarza policji, punktualnie przybył na umówione miejsce. [Zbieg okoliczności, Chmielewska Joanna]
5	Прибыв, сначала позвонил, потом постучал и лишь после этого взялся за ручку входной двери —все, как положено по инструкции. [Стечение обстоятельств, Хмелевская И.]	Najpierw zadzwonił, potem zapukał, a wreszcie przycisnął klamkę, wszystko jak się należy, zgodnie z regułami. [Zbieg okoliczności, Chmielewska Joanna]
6	Прибыв, сначала позвонил, потом постучал и лишь после этого взялся за ручку входной двери —все, как положено по инструкции. [Стечение обстоятельств, Хмелевская И.]	Najpierw zadzwonił, potem zapukał, a wreszcie przycisnął klamkę, wszystko jak się należy, zgodnie z regułami. [Zbieg okoliczności, Chmielewska Joanna]
7	—вежливо спросил сотрудник полиции и вошел в квартиру. [Стечение обстоятельств, Хмелевская И.]	— spytał grzecznie i wszedł do środka. [Zbieg okoliczności, Chmielewska Joanna]
8	Труп, ясное дело, не ответил. [Стечение обстоятельств, Хмелевская И.]	Denat odpowiedzi mu nie udzielił. [Zbieg okoliczności, Chmielewska Joanna]

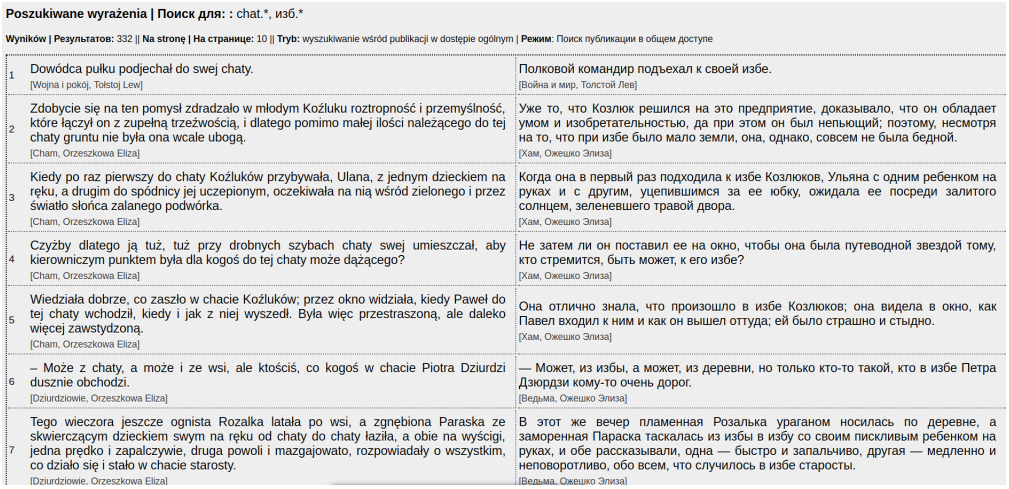
Видно, что поиск справился, однако тексты, где встретилось больше одной словоформы в подходящей форме, выдаются столько раз, сколько в них подходящих словоформ, что не очень хорошо.

Посмотрим на употребление слов «хата» и «изба» и их польских корелятов «chata» и «izba». Сделав запрос в параллельном корпусе. Для этого найдём количество вхождений для каждого варианта из комбинаций русского и польского слова, а также количество вхождений этих лемм отдельно. Приведём пример поиска по одной из лемм и параллельного поиска одной из комбинаций:

Poszukiwane wyrażenie | Поиск для: *chata*

Wyników | Результатов: 1196 || **Na stronę| На странице:** 10 || **Tryb:** wyszukiwanie wśród publikacji w dostępie ogólnym | **Режим**: Поиск публикации в общем доступе

1	Są też w Gradowie domy o nieco przyzwoitszym niż chaty wyglądzie. [Miasto Gradow , Platonow Andrzej]	Есть в Градове жилища и поприличней хат. [Город Градов, Платонов Андрей П.]
2	Za oknem migały chaty jakiegoś miasteczka i bez pośpiechu pomachiwał starymi skrzydłami wiatrak, mieląc z wysiłkiem grube ziarno. [Miasto Gradow , Platonow Andrzej]	За окном проскакивали хижины какого-то городка и не спеша помахивала мельница ветхими крыльями, тяжело меля грубое зерно. [Город Градов, Платонов Андрей П.]
3	Jak stary konował — co to rozpruł krocie kałdunów, co odrąbał mnóstwo kończyn — w kurnych chatach i przy drogach — patrzy na śnieżnobiałą praktykantkę-medyczkę, tak właśnie patrzył Stalin na Titę. [Krąg Pierwszy , Sołżenicyn Aleksander]	Как старый коновал, перепоровший множество этих животов, отсекший несчетно этих конечностей в курных избах, при дорогах, смотрит на беленькую практикантку-медичку, — так смотрел Сталин на Тито. [В круге первом, т. 1, Солженицын Александр]
4	— powiedziała starucha i szurając kaloszami weszła przez niskie drzwi do chaty. [Krąg Pierwszy , Sołżenicyn Aleksander]	— Или приминает она их? — размышляла старуха, шаркая в избу, к низкой двери. [В круге первом, т. 1, Солженицын Александр]
5	Gmerały w nich staruchy, każda na własnym pogorzelisku, coś odkopywały z popiołów, gdzieś chowały, wyobrażając sobie, że są skryte przed ludzkim wzrokiem, jak gdyby wokół nich nadal wznosiły się ściany chat. [Doktor Żywago, Pasternak Borys]	На их поверхности копошились старухи погорелки, каждая на своем собственном пепелище, что-то откапывая в золе и все время куда-то припрятывая, и воображали себя укрытыми от посторонних взоров, точно вокруг них были прежние стены. [Доктор Живаго, Пастернак Борис]
6	— pytał Gordon doktora Żywagę, gdy ten przychodził na obiad do galicyjskiej chaty, w której kwaterowali. [Doktor Żywago, Pasternak Borys]	— спрашивал Гордон доктора Живаго, когда тот приходил днем домой обедать в галицийскую избу, в которой они стояли. [Доктор Живаго, Пастернак Борис]
7	Żona starego wybiegała z pobliskiej chaty na drogę, z krzykiem wyciągała ku niemu ręce i za każdym razem lękliwie się chowała. [Doktor Żywago, Pasternak Borys]	Из-за противоположной избы выбегала на дорогу, с криками протягивала руки к старику и каждый раз вновь боязливо скрывалась его старуха. [Доктор Живаго, Пастернак Борис]



Poszukiwane wyrażenia | Поиск для: : chat.*, изб.*

Wyników | Результатов: 332 || **Na stronę | На странице:** 10 || **Tryb:** wyszukiwanie wśród publikacji w dostępie ogólnym | **Режим:** Поиск публикации в общем доступе

1	Dowódca pułku podjechał do swej chaty. [Wojna i pokój, Tołstoj Lew]	Полковой командир подъехал к своей избе. [Война и мир, Толстой Лев]
2	Zdobycie się na ten pomysł zdradzało w młodym Kożluku roztropność i przemyślność, które łączył on z zupełną trzeźwością, i dlatego pomimo małej ilości należącego do tej chaty gruntu nie była ona wcale ubogą. [Cham, Orzeszkowa Eliza]	Уже то, что Козлюк решился на это предприятие, доказывало, что он обладает умом и изобретательностью, да при этом он был непьющий; поэтому, несмотря на то, что при избе было мало земли, она, однако, совсем не была бедной. [Хам, Ожешко Элиза]
3	Kiedy po raz pierwszy do chaty Kożluków przybywała, Ulana, z jednym dzieckiem na ręku, a drugim do spódnicy jej uczepionym, oczekiwała na nią wśród zielonego i przez światło słońca zalanego podwórka. [Cham, Orzeszkowa Eliza]	Когда она в первый раз подходила к избе Козлюков, Ульяна с одним ребенком на руках и с другим, уцепившимся за ее юбку, ожидала ее посреди залитого солнцем, зеленевшего травой двора. [Хам, Ожешко Элиза]
4	Czyżby dlatego ją tuż, tuż przy drobnych szybach chaty swej umieszczał, aby kierowniczym punktem była dla kogoś do tej chaty może dążącego? [Cham, Orzeszkowa Eliza]	Не затем ли он поставил ее на окно, чтобы она была путеводной звездой тому, кто стремится, быть может, к его избе? [Хам, Ожешко Элиза]
5	Wiedziała dobrze, co zaszło w chacie Kożluków; przez okno widziała, kiedy Paweł do tej chaty wchodził, kiedy i jak z niej wyszedł. Była więc przestraszoną, ale daleko więcej zawstydzoną. [Cham, Orzeszkowa Eliza]	Она отлично знала, что произошло в избе Козлюков; она видела в окно, как Павел входил к ним и как он вышел оттуда; ей было страшно и стыдно. [Хам, Ожешко Элиза]
6	– Może z chaty, a może i ze wsi, ale ktościś, co kogoś w chacie Piotra Dziurdzi dusznie obchodzi. [Dziurdziowie, Orzeszkowa Eliza]	— Может, из избы, а может, из деревни, но только кто-то такой, кто в избе Петра Дзюрдзи кому-то очень дорог. [Ведьма, Ожешко Элиза]
7	Tego wieczora jeszcze ognista Rozalka latała po wsi, a zgnębiona Paraska ze skwierczącym dzieckiem swym na ręku od chaty do chaty łaziła, a obie na wyścigi, jedna prędko i zapalczywie, druga powoli i mazgajowato, rozpowiadały o wszystkim, co działo się i stało w chacie starosty. [Dziurdziowie, Orzeszkowa Eliza]	В этот же вечер пламенная Розалька ураганом носилась по деревне, а замоpенная Параска таскалась из избы в избу со своим писклявым ребенком на руках, и обе рассказывали, одна — быстро и запальчиво, другая — медленно и неповоротливо, обо всем, что случилось в избе старосты. [Ведьма, Ожешко Элиза]

При этом поиск по параллельному корпусу выполняется по основе слова (например, «изб.*»), чтобы найти все формы (как мы помним, поиск по лемме нельзя использовать). Интересно, что при непараллельном поиске было бы много побочных результатов, которые начинаются на «изб», например «избежать», однако из-за параллельного поиска результаты получаются достаточно точными во всех случаях, кроме «изб.* izb.*» (в последнем случае приходится перебирать все словоформы данных лемм).

изба	1256
хата	741
izba	1595
chata	1196
изб.* chat.*	332
изба izba[2]	152
хат.* chat.*	235
хат.* izb.*	16

Полученные частотности можно увидеть в таблице слева. Бросающееся в глаза несоответствие количества лемм и количества их сочетаний объясняется, во-первых, наличием у польских слов нескольких значений, не все из которых соответствуют русским (например, «izba» может значить «комната»), во-вторых, другими словами, выражающими интересующие нас значения, такими как польское «chałupa». Проанализировав частотности, мы получаем достаточно интересные результаты. Во-первых, слово «хата» в целом реже используется в русском языке. Это логично, ведь это слово южнорусских говоров, а не лёгших в основу литературного языка среднерусских. Во-вторых, видно заметное влияние языка оригинала при переводе на другой: ситуация, когда у соответствующих слов один и тот же (исторически) корень встретилась примерно одинаковое количество раз и для «изб», и для «хат». В-третьих, для выражения интересующего нас значения в польском предпочитается слово «chata» (в русском, как мы знаем «изба»): соответствие «хат.* izb.*» очень мало, по сравнению с соотношением «изб.* chat.*». При этом, скорее всего, среди

[2] О запросах для поиска данного сочетания лемм см. выше.

случаев, когда выбираются слова с разными корнями, будут значительно чаще встречаться тексты, переведённые с третьего языка на русский и польский, однако этот вопрос требует дополнительного исследования.

Приложение

	количество
Польские художественные тексты и их переводы на русский язык	45
Русские художественные тексты и их переводы на польский язык	34
Польская пресса и ее переводы на русский язык	14
Русская пресса и ее переводы на польский язык	28
Польские и русские религиозные тексты „Библия тысячелетия" и синодальный перевод Библии	76
Польские юридические тексты и их переводы на русский язык	2
Русские юридические тексты и их переводы на польский язык	0
Русские и польские переводы иностранных художественных текстов	8