Марек Лазинский
Магдалена Куратчик
*Польша, Варшавский университет*

# Польско-русский параллельный корпус Варшавского университета

**Ключевые слова:** корпусная лингвистика, корпус параллельных текстов, польский и русский языки
**Key words:** corpus linguistics, parallel corpus, Polish, Russian

## Abstract

The Polish-Russian Parallel Corpus has been developed at the University of Warsaw (the Faculty of Polish Studies and the Institute of Russian Studies) in co-operation with the National Corpus of Polish and the Russian National Corpus. The annotation and search possibilities in the corpus result from the annotation of the co-operating national corpora. The search interface is based on the user-friendly interface of the Russian National Corpus. The corpus consists of Russian and Polish literary classics (90%), nonfiction books and legal texts (5%), religious texts (i.e. Bible translations; 4%) and contemporary press articles (1%). Great Russian realistic novels of the 19th century, together with the modern Russian books (e.g. by Alexandr Solzhenitsyn and Victor Erofeyev) which are the most popular in Poland, made up a significant part of the corpus. We have also taken into account these works of Polish literature that are the most widely known in Russia. Looking for *loci communes* in the Russian and Polish cultures was an important, extra-linguistic aspect of the corpus project.

Польско-русский корпус Варшавского университета (далее П-РК) – это представительный (т. е. со сбалансированным составом текстов), аннотированный (снабженный морфосинтаксической и библиографической разметкой), лишенный омонимии параллельный корпус. Как и другие параллельные корпуса, П-РК может оказать практическую помощь переводчикам, лингвистам, лексикографам, исследователям литературы и культуры.

## История проекта

П-РК – первый созданный в Польше общедоступный параллельный корпус. Хотя параллельные тексты составляются в рамках различных проектов, например европейских – MetNet и Clarin, однако они пока недоступны широкому кругу пользователей.

Мысль создать П-РК зародилась в 2009 г. во время работы над Польским национальным корпусом (далее – ПНК). С идеей сотрудничества выступили полонисты и русисты из Уфы – Елена Слободян и Борис Орехов. С польской стороны реализацией проекта заинтересовались полонисты и русисты Варшавского университета: Марек Лазинский из Института польского языка, Магдалена Куратчик из Института русистики и Наталья Годлевская – аспирантка Института польского языка. В таком составе в 2010 г. – под руководством Марека Лазинского – группа приступила к реализации гранта Национального центра науки NN104056638. Существенную помощь в реализации проекта, также при выравнивании текстов, оказали сотрудники Национального корпуса русского языка (НКРЯ) – Дмитрий Сичинава и Светлана Минлос. При подготовке и выравнивании текстов принимали участие и другие разработчики, менялась также техническая поддержка проекта.

ПНК создавался в рамках соглашения между Институтом основ информатики, Институтом польского языка Польской академии наук, кафедрой компьютерной и корпусной лингвистики Лодзинского университета и Научным издательством PWN при участии лингвистов Варшавского университета. Объем корпуса составляет 1,5 млрд. слов, в том числе 300 млн. в сбалансированном подкорпусе (NKJP 2012).

Для разметки текстов и снятия омонимии П-РК использует морфологический анализатор TAKIPI. Принципы и проблемы проекта еще перед его окончанием были описаны в [Łaziński et al. 2012].

## Структура корпуса

Совокупный объем П-РК достигает 30 миллионов словоупотреблений наполовину польских, наполовину русских. Поскольку достижение образцовой структуры корпуса, охватывающей равное количество слов в исходных польских и русских текстах, оказалось трудным в осуществлении, мы решились на перевес польских текстов над русскими в соотношении три к двум. В результате база текстов содержит 50% польских оригиналов, 33% русских и 15% переводов с третьего языка, в том числе переводы Библии и тексты международных договоров. Эта статистика касается, конечно, лишь языка

оригинала, поскольку целевой корпус содержит, по всем правилам, равное количество русских и польских слов (в нашем случае с незначительным перевесом русских).

Проект корпуса предполагал включение значительного количества художественных текстов, в том числе классики, которая не подлежит охране авторским правом, и может использоваться без разрешения. Стоит заметить, что художественные тексты составляют основную часть различных параллельных корпусов. Также в нашем проекте удельный вес художественной литературы достиг 90%. Мы задались, однако, целью обогатить корпус за счет публицистики, включая переводы статей журнала «Forum» и польских газетных текстов, переводы которых публикуются на сайте inosmi.ru. Несмотря на то, что эти тексты составляют лишь один процент всего корпуса, они являются незаменимым источником материала при исследовании современных нехудожественных текстов. Включение публицистики выделяет наш корпус среди существующих параллельных корпусов. Так, например, проект Intercorp Чешского национального корпуса хотя и содержит газетные тексты PressEurope (presseurop.eu), однако тематически эти тексты ограничиваются проблематикой Евросоюза (кроме того, PressEurope не включает русских текстов).

П-РК содержит также 4% религиозных текстов. К ним относятся наиболее известные в обеих культурах переводы Нового завета и части Ветхого, т. е. Библия Тысячелетия и русский Синодальный перевод. Несмотря на изобилие библейских сайтов, снабженных конкордансерами, ни в одном из них нет возможности одновременного просмотра русского и польского текстов.

Остальные 5% объема корпуса составляют юридические тексты и документальная литература.

Около 66% текстов находится в открытом доступе. Остальные тексты, являющиеся объектом авторского права, – в закрытом, и доступны лишь некоторым сотрудникам Варшавского университета. Проблемы с получением согласия на использование текстов для корпусных целей (или хотя бы ответа со стороны обладателей авторских прав на просьбы сотрудников проекта) касались не только современных текстов, но и русской классики. Стоит заметить, что романы Ф. Достоевского и Л. Толстого долго не переводились, так как на значительной части польской территории официальным языком был русский. Большая часть классики была переведена только в 20–30 гг. XX в. молодыми литераторами, которые умерли в 60–70 гг., и авторские права к этим переводам еще долго будут защищаться.

## Культурный аспект корпуса

Наш проект преследует не только практические и научные, но и культурные цели. Мы стремились собрать тексты русских авторов, особенно популярных в Польше и важных для польской картины России, а также тексты польских авторов, популярных в России. Не случайно собирать русские тексты мы начали с переводов А. Солженицына, а первое согласие на использование своих переводов мы получили от знаменитого русиста Ежи Помяновского. Из современных русских писателей невозможно было проигнорировать тексты Виктора Ерофеева, активного участника дебатов по насущным вопросам польско-русских отношений. Межкультурные контакты отражены также в газетных текстах, подобранных редакциями журнала «Forum» и сайта inosmi.ru как затрагивающие вопросы, представляющие собой интерес для носителей обеих культур.

Культурному сближению способствует также библиографическое описание художественных текстов, выходящее за рамки стандартной аннотации. Подготовленные сотрудниками Кафедры русской литературы короткие описания русских художественных текстов представляют собой краткий путеводитель по русской культуре.

Польские сотрудники проекта также смогли узнать что-то новое о специфике восприятия польской культуры в России. К нашему удивлению, первым текстом, о котором попросили нас русские партнеры, был *Фараон* Б. Пруса – свободный от авторских прав. С точки зрения поляков *Фараон* не самый важный роман Б. Пруса, его давно уже нет в списке школьной литературы и, возможно, читают его лишь только студенты-полонисты. Выбор именно *Фараона* можно, конечно, объяснить личными симпатиями русских коллег, однако не менее вероятной причиной может быть удивительная популярность в России этого универсального романа о власти[1]. Поисковая система Google находит в русскоязычном поиске 23 000 ссылки на *Фараона* и лишь 12 000 – на *Куклу*, в то время как польскоязычный Google раскрывает обратную пропорцию: *Lalka* – 173 000, *Faraon* – 51 000. Работа с корпусом, оказывается, учит не только конечного пользователя.

[1] Отметим, не придавая этому факту особого значения, что *Фараон* был одним из любимых романов Сталина (J. Czapski, *Na nieludzkiej ziemi*). Такую же информацию приводит русская Википедия.

## Разметка текстов и поисковый аппарат

Тексты корпуса хранятся в реляционной базе данных[2]. Ячейки таблиц базы содержат информацию о словах, предложениях, текстах, а также о соотношении предложений в выровненном тексте. Язык запросов SQL обеспечивает поиск нужных элементов, и – в отличие от корпусов, осуществляющих поиск индексированных текстов, например, ПНК, – пользователь может не знать системы метатекстовой разметки. Для любого запрашиваемого слова вместе с присвоенными ему морфосинтаксическими метками (тэгами) определяется его место в предложении, и к этому предложению программа подыскивает эквивалент на другом языке.

При выравнивании текстов была использована программа ABBY Aligner. Польские тексты размечались морфологизатором TAKIPI, используемым в свое время ПНК (актуальная версия ПНК размечена новым морфологизатором – Pantera). Русские тексты размечались морфологизатором Mystem.

Поисковая система предоставляет возможность искать в исходном языке слова и сочетания слов – с использованием языка поисковых запросов: дизъюнкции (или то, или это, или оба сразу), конъюнкции, отрицания – и выдает эквивалентные им слова и предложения в тексте перевода. Поиск может осуществляться также по любым морфологическим признакам. При этом используются разметки национальных корпусов – польского и русского. Так, возможен поиск заданных словоформ, например, глаголов в форме инфинитива, глаголов в форме множественного числа повелительного наклонения, или заданных лексем, к примеру, форм глагола *iść*. В последнем случае достаточно вписать в окно поиска словарную форму, чтобы найти все словоформы, в частности: *idę*, *idziecie*, *szedłby*, *szłyśmy*.

## Принципы поиска

Оконный интерфейс поиска напоминает интерфейс Национального корпуса русского языка. Он разделен на поиск точных форм и лексико-грамматический поиск. Интерфейс, при котором достаточно поставить «галочку» при названии нужной категории или грамматического класса, намного удобнее, чем формализованный язык запросов ПНК и большинства других национальных корпусов, создаваемых для пользователей-специалистов. Приспособить систему категорий ПНК к традиционному делению слов на части речи было

[2] Авторами программы и администраторами базы являются Павел Годлевский и Кшиштоф Осецкий.

трудно, однако благодаря успеху дела даже ученик гимназии может осуществлять поиск в нашем корпусе.

Как известно, состав русских и польских грамматических категорий не совпадает. К тому же, морфологизатор для польского языка построен на флексемах – грамматических классах, намного более подробных, чем школьные. Флексемы представляют собой классы слов с одинаковым набором категорий. Так, например, форма непрошедшего времени польского глагола спрягается по лицам и числам, глагол в повелительном наклонении имеет лишь формы первого и второго лица обоих чисел, а инфинитив – неизменяемая форма. В результате названные формы образуют три отдельные флексемы.

Таким образом, традиционная категория глагола распадается на 15 разных флексем: непрошедшая форма – *fin*, будущая простая форма *będzie* – *bedzie*, аглютинант[3] *być* – *aglt*, псевдопричастие[4] – *praet*, императив – *impt*, имперсональная форма – *imps*, инфинитив – *inf*, деепричастие предшествования (прошедшего времени) – *pant*, деепричастие одновременности (настоящего времени) – *pcon*, действительное причастие (несовершенного вида) – *pact*, страдательное причастие – *ppas*, герундий (формы *-nie*/*-cie*) – *ger*, глаголы типа *winien* – *winien*, предикатив, например, *warto* – *pred*.

Непосвященному пользователю ПНК значительную трудность доставляет интерпретация форм прошедшего времени. Они всегда состоят из причастия прошедшего времени и показателя лица и числа, который при разметке интерпретируется как отдельное слово, например: zrobił+em; где *zrobił* – форма претерита (флексема *praet*) в ед. числе мужск. р., а *-em* – это аглютинант (*aglt*) 1 л. ед. числа (формы 3 лица употребляются без аглютинантов).

Представленная интерпретация подобных форм, хотя она и не противоречит ни истории польского языка, ни славянской морфологии, приводит к тому, что нельзя задать поиск любых глаголов в форме 1 лица прошедшего времени. Однако, например, поиск любого существительного в единственном числе осуществляется без препятствий. Дело в том, что сведения о числе содержатся в тэге, определяющем имя существительное, в то время как сведения о лице включаются не в тэг *praet*, а тэг аглютинант. Запрос, касающийся форм прошедшего времени, может быть, конечно, осуществлен, однако его реализация отнимает больше времени, так как программа должна найти последовательность двух словоформ.

Как было сказано выше, основной поиск осуществляется в обоих языках с опорой на школьные, традиционно выделяемые категории и классы слов. Поиск по флексемам и категориям морфологизатора TAKIPI возможен при

---

[3] Под аглютинантами понимаются подвижные окончания глагольных форм, ср. *spał**eś** dobrze* и *dobrz**eś** spał* (=(ты) хорошо спал).

[4] *Псевдопричастием* называют личную форму, используемую при образовании форм прошедшего и будущего времени, например: *pracował-eś*, *będziesz pracował*.

включении дополнительных критериев поиска. Традиционные части речи в обоих языках не совпадают, однако довольно близки друг другу. В польскоязычных текстах не различаются неизменяемые части речи за исключением наречия и предлога, которые автоматически узнаются в контексте. Зато в русскоязычных текстах возможен поиск по частицам и междометиям.

Категория «падеж» в польской части содержит окно «звательный», а в русской части выделен «партитив». Только русская часть содержит окна для поиска имени прилагательного по краткой/полной форме и существительного по одушевленности/неодушевленности, и только польская дает доступ к поиску имперсональных форм (на *-no*/*-to*).

Как и все двуязычные корпуса, П-РК не свободен от ошибок, связанных, в частности, с автоматическим выравниванием текстов. Может случиться, что эквивалент запрашиваемого слова отсутствует в переводе лишь потому, что программа поиска выдает в исходном языке более широкий контекст, чем в переводе.

При разработке системы поиска и базы текстов более важной целью было для нас создать удобный интерфейс, чем построить сложную систему многоаспектного поиска. Поэтому, несмотря на упомянутые достоинства, наша программа страдает некоторыми явными недостатками, часть из которых можно исправить в будущем, если, конечно, проект будет продолжаться. Оконный интерфейс не дает, например, возможности осуществлять поиск с использованием разметки XML. Интерфейс морфологического поиска не позволяет строить запросы, касающиеся сочетаний слов во всех возможных их формах. Можно, к примеру, задать поиск любого существительного женского рода в дат. п. мн. числа, но нет возможности построить запрос, касающийся сочетания глагола с таким существительным. Словосочетание можно найти лишь в какой-либо конкретной грамматической форме его компонентов с использованием для этого окна поиска слов. Зато поиск слов дает возможность параллельного поиска, при котором тексты обоих языков должны отвечать заданным требованиям.

В качестве иллюстрации приведем следующий пример. Будем искать такие русские контексты, в которых появляется представляющее трудность многим полякам прилагательное *российский* (набранное с использованием регулярных выражений Российск.*|российск.*), причем в эквивалентных им польских текстах не появляется прилагательное *rosyjski* (rosyjs.*|Rosyjs.*). Программа выдает 32 таких примера, составляющих 15% от 195 всех случаев употребления прилагательного *российский*. Аналогичный поиск – на этот раз касающийся прилагательного *русский* – выдает 650 примеров, составляющих 37% от 1757 употреблений данного прилагательного. Оказывается, что прилагательное *российский* (употребляемое как согласованное определение) чаще, чем *русский*, переводится существительным в родительном падеже – *Rosji* (т. е. несогласованным определением). Довольно часто, однако, оно

или оставляется без перевода, или – в 85% случаев – передается прилагательным *rosyjski*. Прилагательное *русский* чаще, чем *российский*, получает переводной эквивалент (каждый третий пример употребления), поскольку при описании древности и в текстах, стилизованных под разговорную речь, ему соответствует прилагательное *ruski*, и, нельзя забывать, что *русский* – это по-польски просто *Rosjanin*.

## Разметка текстов и снятие омонимии

Тексты обоих языков размечены таким образом, что каждой словоформе присваивается соответствующая лемма, и определяются ее грамматические категории. В польских текстах дополнительно снята омонимия. Русский морфологизатор Mystem не снимает омонимию (в НКРЯ она снята частично, причем вручную). Морфологический поиск очень часто выдает результаты, содержащие некоторое количество «шума», т. е. формы с одинаковым написанием получают весь возможный набор разборов. В нашем корпусе поиск форм польского существительного *dama* выдаст, например, контекст: *odciski palców obu dam*, но не: *Dam panu znać*. Однако поиск русского существительного *дама* выдаст в качестве результата как *обеих дам*, так и *Я вам дам знать*.

При поиске форм заданных лексем, например *dama*, омонимия не представляет собой проблемы, так как ошибочные результаты легко исключить вручную. Зато при поиске любых форм, обладающих заданными характеристиками, регулярный синкретизм является непреодолимым препятствием. Так, в нашем корпусе мы можем искать любые прилагательные мужского или среднего рода с окончанием *-ym*/*-im*, и (если не принимать во внимание ошибки программы по снятию омонимии) мы получим лишь формы творительного падежа, а не омонимичные им формы предложного. При поиске по русскому тексту нет смысла задавать поиск аналогичного типа (например, прилагательных женского рода с окончаниями *-ой*/*-ей*), так как среди результатов получим не только формы творительного падежа, но и предложного.

## Нынешнее состояние и перспективы развития

П-РК представляет собой закрытый проект. Он будет, конечно, продолжать свое функционирование на сервере Варшавского университета, однако его совершенствование и пополнение потребовало бы новых затрат. Корпус мог бы быть, например, интегрирован с другими корпусами в рамках большого

многоязычного проекта, включающего польский язык. Такого типа проекты уже реализуются (в частности, Intercorp Parasol), однако пока не в Польше. Естественной платформой для осуществления такой идеи мог бы стать Национальный корпус польского языка, но он также является закрытым проектом.

П-РК в его актуальном состоянии отвечает основным требованиям переводчиков и лингвистов и, как кажется, показывает польскую лингвистику в неплохом свете. То обстоятельство, что первый польский общедоступный двуязычный корпус создан в рамках международного сотрудничества университетов и национальных корпусов, противоречит распространенному мнению о значительном снижении интереса к русскому языку и культуре в Польше и польскому в России.

## Литература

Łaziński M., Kuratczyk M., Orekhov B., Słobodjan E., 2012, The Polish-Russian Parallel Corpus and Its Application in the Linguistic Analysis, *Prace Filologiczne* LXIII, c. 209–218.

*Narodowy Korpus Języka Polskiego*, 2012, A. Przepiórkowski, M. Bańko, R. Górski, B. Lewandowska-Tomaszczyk (red.), Warszawa: Wydawnictwo Naukowe PWN.