О. А. Волчек, В. В. Порицкий
O. A. Volchek, V.V. Poritski

ЭКСПЕРИМЕНТАЛЬНЫЙ КОРПУС БЕЛОРУССКОГО ЯЗЫКА: ТЕКУЩЕЕ СОСТОЯНИЕ И ПЕРСПЕКТИВЫ РАЗВИТИЯ

AN EXPERIMENTAL CORPUS OF BELARUSIAN: ITS PRESENT AND FUTURE

Аннотация. В 2012 г. в БГУ была начата работа над корпусом газетных и художественных текстов на белорусском языке. Сейчас объем текстовой базы корпуса приближается к 9 млн с/у. В статье описаны состав и структура ресурса, процедура подготовки текстов и реализация поискового механизма. Приводится пример использования корпуса в лингвистических исследованиях.

Abstract. An experimental corpus of Belarusian fiction and newspaper articles has been under development at the BSU since 2012. Current size of the corpus is slightly less than 9 mln tokens. This paper accounts for an intermediate stage of the project: we describe the corpus' content and structure, present the accompanying search tool and discuss some text preprocessing issues. We also exemplify one possible use of the corpus in linguistic studies.

1. Введение

Свои национальные корпуса имеют многие славянские языки, в том числе русский (**http://ruscorpora.ru**), польский (**http://nkjp.pl**) и украинский (**http://lcorp.ulif.org.ua/virt_unlc**). В Беларуси ситуация обстоит существенно хуже: корпусная лингвистика рассредоточена по нескольким научным центрам, ни один из которых пока не занимается столь масштабным проектом. Коллектив исследователей из ГрГУ, возглавляемый Л. В. Рычковой, работает над корпусом газет Гродненской области. В МГЛУ под руководством А. В. Зубова создаются небольшие параллельные корпуса, доступ к которым, впрочем,

ограничен. В 2011 г. в сети был размещен «Corpus Albaruthenicum» (**http://grid.bntu.by/corpus**), разработанный в Институте языка и литературы НАН РБ в основном усилиями В. А. Кощенко. Этот ресурс содержит 350 тыс. с/у (74 научных текста) с морфологической разметкой. Наконец, совсем недавно НКРЯ был обогащен белорусско-русским и русско-белорусским параллельными подкорпусами общим объемом около 3 млн с/у (**http://ruscorpora.ru/search-para-be.html**). В них представлена преимущественно художественная литература; они оснащены удобным поисковым интерфейсом и могут быть полезны филологам-славистам и изучающим белорусский язык. К сожалению, доступ к полным текстовым базам параллельных подкорпусов НКРЯ и «Corpus Albaruthenicum» закрыт.

Как видно, для белорусского языка остро стоит проблема создания представительного морфологически аннотированного одноязычного корпуса, который бы свободно распространялся и был пригоден для оффлайнового использования. В нашей статье излагаются промежуточные итоги работы над экспериментальным корпусом, призванным соответствовать этим требованиям. Текст представляет собой расширенный вариант нашей более ранней заметки[1].

2. Состав и структура корпуса

На сегодняшний день в состав корпуса входят подкорпуса газетных текстов и художественной прозы. Газетный подкорпус содержит издания двух периодов:

1) середина XX в. Это годовые архивы газет «Голас Радзімы» (май 1961 – апрель 1962 гг.) и «Літаратура і мастацтва» (январь–декабрь 1961 гг.) общим объемом около 2 млн с/у.

2) начало XXI в. Это годовые архивы газет «Звязда» и «Чырвоная змена» (август 2008 – июль 2009 гг.), а также двухго-

[1] *Волчек О. А., Порицкий В. В.* Проект корпуса белорусскоязычной периодики и художественной прозы // Компьютерная лингвистика: научное направление и учебная дисциплина. 2012. Вып. 2. С. 16–19.

дичный архив газеты «Голас Радзімы» (2008–2009 гг.) – всего около 4,4 млн с/у.

Состав подкорпуса художественной прозы таков:

1) подборка текстов первой половины XX в. На данный момент в нее вошла только проза Я. Коласа (248 тыс. с/у). В ближайшее время мы не планируем существенно расширять эту текстовую базу, потому что представительная выборка белорусской прозы XX в. уже содержится в параллельном подкорпусе НКРЯ.

2) современная белорусская литература. Источником материала стали журналы «Полымя», «Маладосць» и «Дзеяслоў» за 2009–2011 гг. Из каждого номера отбирались лишь тексты, отнесенные редакцией к рубрике «Проза». Размер этой части подкорпуса – примерно 2,2 млн с/у (около 200 разных авторов).

С технической стороны корпус представляет собой иерархию папок. Каждый текст хранится в отдельном текстовом файле с кодировкой ANSI 1251. Художественная проза дополнительно снабжена метатекстовой информацией: автор, заглавие, жанровая принадлежность. Подробнее ознакомиться с устройством ресурса можно на странице **https://github.com/poritski/YABC**, где размещен его документированный демонстрационный фрагмент.

3. Сбор и подготовка текстов

Публикации газет «Звязда», «Чырвоная змена», «Голас Радзімы» и журнала «Дзеяслоў» размещены в Интернете в виде HTML-страниц. Для их получения было разработано два скрипта на языке PHP, первый из которых формирует списки требуемых страниц, а второй скачивает эти страницы. При очистке данной части корпусного материала от метатекстовой разметки, как всегда, возник ряд трудностей, связанных с выделением собственно текстового фрагмента, неунифицированной передачей отдельных символов (тире, кавычек, белорусского і), спецсимволами HTML и др.

Чтобы справиться с этими сложностями, был написан простой скрипт очистки на языке Perl. На целевом множестве файлов наш скрипт обеспечивает качество подготовки текста, близкое к

100%, но при этом не является универсальным, т.е. не может эффективно очищать от метатекстовой разметки, например, блоговые записи.

Особая проблема – повторяющиеся фрагменты текста внутри одной статьи. Разработан скрипт, способный с хорошей точностью идентифицировать такие дубли, но окончательное решение об удалении мы оставляем за пользователем.

Произведения, опубликованные в «Полымі» и «Маладосці», были извлечены из PDF-файлов соответствующих номеров журналов. При этом не обошлось без двух затруднений. Во-первых, конвертер PDF→TXT распознает журнальную страницу, сверстанную в несколько колонок, как одну колонку сплошного текста с группами пробелов посередине; дробить текст на колонки по группам пробелов переменной длины технически неудобно. Во-вторых, проблематично разграничить знак переноса и дефис в конце графической строки. Все это потребовало тщательной вычитки сконвертированного текста.

Материал для диахронной части газетного подкорпуса извлекался из PDF-файлов подшивок, оцифрованных в Национальной библиотеке Беларуси. Качество текстов, полученных в результате распознавания с помощью программы ABBYY FineReader, сильно различается и зависит, в первую очередь, от качества исходных скан-копий. Хотя «Літаратура і мастацтва» распознается существенно лучше, чем «Голас Радзімы», в материале из обеих газет, вдобавок к уже отмеченным трудностям (смешение текста из разных колонок, неразличение переносов и дефисов), появляется новая – регулярные ошибки распознавания. Чаще всего это пропуски символов и их замены (*а* → *в*, *д* → *л*, *з* → *э*, *й* → *іі*, *ы* → *ьі* и др.).

Чтобы решить эту проблему, мы использовали Грамматический словарь белорусского языка, электронный вариант которого доступен на сайте **http://slounik.org**. В несколько проходов по текстам был сформирован список подстановок для значительной части словоформ, которые не обнаружились в словаре. Это:

1. Словоформы с лишним дефисом – знаком переноса.

2. Самые высокочастотные (первые несколько тысяч) «неопознанные» словоформы, содержащие явные ошибки распознавания. Для них подстановки выписывались вручную.

3. Достаточно длинные низкочастотные «неопознанные» словоформы, которые находятся на единичном расстоянии в метрике Левенштейна от однозначно идентифицируемых словоформ из словаря.

Если словоформа отсутствует в словаре и списке подстановок, к ней применяется последовательность регулярных выражений, которые в типичных контекстных условиях заменяют ошибочный символ на верный (*...аай* → *...вай*, *...іэа...* → *...іза...*, *...лыі...* → *...льн...* и др.).

4. Поисковый механизм

В нынешней реализации ресурса к каждому из подкорпусов прилагаются две программы на языке Perl: индексатор и основной поисковый скрипт. Они обеспечивают быстрый поиск точных форм слов, в том числе любых наборов форм (например, парадигм склонения), и вывод найденных контекстов в файл.

Скрипт-индексатор последовательно читает все файлы подкорпуса, токенизирует их (при этом используется адаптированный токенизатор Г. Шмида из проекта TreeTagger[2]) и строит обратный индекс – таблицу соответствий между словоформами и их позициями. В токенизированных текстах каждая словоформа размещена на отдельной строке, поэтому обратный индекс «запоминает» не только имена файлов, но и номера строк.

В служебном файле *wordlist.txt* пользователь задает поисковые запросы. Каждый запрос записывается на одной строке и состоит из регулярного выражения и его идентификатора, разделенных знаком табуляции. В частности, регулярное выражение мо-

---

[2] *H. Schmid.* Probabilistic part-of-speech tagging using decision trees // Proceedings of international conference on new methods in language processing. Manchester, 1992.

жет определять одну словоизменительную парадигму, а идентификатор – совпадать с начальной формой слова. Пример запроса:

небасхіл(а(ў|м(і|і)?|х)?|ы|у|е)? небасхіл

Основной поисковый скрипт, обращаясь к индексу и текстовой базе, обрабатывает запросы. Результаты поиска скрипт выводит в отдельный текстовый документ, где каждому вхождению соответствует одна строка, содержащая информацию о файле-источнике, идентификатор и точную найденную форму в окружении левого и правого контекстов, ширину которых пользователь может задать самостоятельно. Таким образом, наш поисковый механизм представляет собой простейший корпусный менеджер, пока не оснащенный графическим интерфейсом.

5. Пример использования

Наш экспериментальный корпус уже позволяет получать лингвистические результаты, которых трудно добиться с помощью других существующих ресурсов. Вот пример одного такого наблюдения, относящегося к лексической семантике.

Мы подготовили запрос для слова *небасхіл* и проанализировали выдачу, полученную по всем подкорпусам. Как видно из табл. 1, употребление этой лексемы свойственно больше художественной литературе, чем газетному тексту.

Таблица 1. Частота употребления лексемы *небасхіл* (на 1 млн с/у)

Худ. литература		**Периодика**	
Я. Колас	Нач. XXI в.	Сер. XX в.	Нач. XXI в.
20,2	12,1	4,0	4,8

За последние 100 лет прозаики стали использовать слово *небасхіл* реже и сделали его семантику более размытой. Если употребление этого существительного в текстах Я. Коласа реализует

те и только те лексические значения, которые описаны в словаре[3], то сегодня литераторы начинают называть *небасхілам* не только горизонт или часть неба над ним, но и небо целиком: *Высока ў небе спяваў жаўранак. Чорная кропка ў бязмежным сінім сусвеце. Дрыготкая, яна ўзнімалася ўсё вышэй, пакуль плаўна не растварылася ў небасхіле* (Маладосць. 2009. № 3).

В периодике 1960-х гг. *небасхіл* выступает преимущественно как синоним к слову *гарызонт*. В языке современных газет, как и в художественной прозе, семантика этого слова расширилась за счет контекстов, в которых существительным *небасхіл* обозначается все небо. Кроме того, метафора «Талантливые люди или культурные ценности – это звезды», вероятно, унаследованная из русского языка, повлияла на развитие у слова *небасхіл* регулярного переносного значения: *Нясвіж – беларуская зорка на еўрапейскім небасхіле* (Звязда. 2009. № 89).

6. Перспективы развития

Первоочередная задача при совершенствовании нашего корпуса – нанесение морфологической разметки. Белорусская машинная морфология, созданная под руководством И. В. Совпеля[4], не распространяется открыто, а ее вариант, разработанный Ю. С. Гецевичем с соавторами для системы NooJ[5], имеет чрезвычайно низкое быстродействие. Лексико-грамматическая база В. А. Кощенко, которая по сути представляет собой высококачественный правиловый лемматизатор, еще не доведена до первого открытого релиза. Мы начали работу над созданием простейшего

---

[3] Тлумачальны слоўнік беларускай мовы: у 5 т. Пад агул. рэд. К. К. Атраховіча. Мінск, 1979. Т. 3. С. 352.

[4] *Совпель И. В., Рубашко Н. К., Невмержицкая Г. П.* Компьютерный фонд белорусского языка и его приложения // Информационные системы и технологии (IST 2006). Минск, 2006. Ч. 2. С. 71–76.

[5] *Hetsevich Yu. S., Hetsevich S. A., Lobanov B. M.* Belarusian and Russian linguistic processing modules for the system NooJ as applied to text-to-speech synthesis. **http://www.dialog-21.ru/digests/dialog2012/materials/pdf/136.pdf**

лемматизатора на основе Грамматического словаря. Этим инструментом, по-видимому, и будет проаннотирован наш корпус.

Еще одна немаловажная задача – совершенствование текстовой базы: ее расширение, улучшение балансировки и повышение качества (устранение опечаток и ошибок распознавания, очистка от дублей).

Наконец, предполагается доработать программную часть, т. к. существующий поисковый механизм не справляется с более чем однословными запросами и не способен учитывать грамматические сведения. В долгосрочной перспективе мы также рассчитываем создать графический интерфейс для локальной работы с корпусом.