Параллельный корпус русских летописей в Интернете: цели, задачи, технологическая основа, использование

A parallel corpus of Russian chronicles on the Internet: goals, challenges, technology, and applications

Аникина Р. А., аспирант, anikina.regina@gmail.com;
Баранов В. А., д-р филол. наук, профессор, зав. кафедрой "Лингвистика", victor.a.baranov@gmail.com
Ижевский государственный технический университет им. М. Т. Калашникова
Ижевск, Россия

Ключевые слова: русские летописи, полнотекстовая база данных, параллельный корпус

Summary

This work shows the goals and challenges for creating a parallel corpus of five of the oldest copies of Russian chronicles as part of the Manuscript project (http://manuscripts.ru). It draws attention to the idiosyncrasies of the corpus, which includes a few types of markup, and to a few data visualization requirements for research on manuscripts using linguo-textological and corpus research methods.

В последние годы все более активно ведутся работы по созданию полнотекстовых машиночитаемых веб-публикаций древнейших и средневековых письменных памятников. Одновременно разрабатываются специализированные программно-инструментальные средства редактирования, обработки, разметки и визуализации таких ресурсов, а также инструменты для использования электронных копий в исторических, лингвистических, текстологических и иных исследованиях.

К сегодняшнему дню создано и доступно в Интернете достаточно много ресурсов на основе русских летописей – как списков Повести временных лет, так и полных транскрипций той или иной рукописи. Например, хорошо известны параллельная публикация нескольких списков ПВЛ, подготовленная Дональдом Островским (http://hudce7.harvard.edu/~ostrowski/pvl/), и критическое издание ПВЛ, осуществленное Дэйвидом Бирнбаумом (http://clover.slavic.pitt.edu/pvl/), параллельный корпус двух списков ПВЛ Института русской литературы (http://www.pushkinskijdom.ru/Default.aspx?tabid=4869) и транскрипции летописей на сайте "Изборник" (http://litopys.org.ua), Полное собрание русских летописей проекта "Рукописные памятники Древней Руси" (http://www.lrc-lib.ru/index.php%3Fid=5) и некоторые другие. Большая часть ресурсов снабжена средствами навигации, которые позволяют перейти к нужной странице или погодной записи. Среди всех

электронных изданий, представляющих собой публикации переводов или сканированных печатных изданий, а также электронные наборы текстов по печатным изданиям. своей полнотой и наличием лингвистической разметки некоторых списков выделяется последняя.

Одновременно с этим к настоящему времени не создано ни одного интернет-ресурса на основе русских летописей, который бы имел и текстологическую, и лингвистическую разметки, позволял бы осуществлять не только стандартный поиск нужного листа или погодной записи, но и выборку по нескольким текстологическим, аналитическим и лингвистическим характеристикам, строить перечни фрагментов, визуализировать отношения между ними в списках, получать упорядоченные списки словоформ и/или лемм и иметь статистические сведения о лингвистических единицах рукописей. Можно сказать, что, несмотря на значительное количество электронных изданий русских летописей в Интернете, все они представляют собой ресурсы, лишь по форме отличающиеся от печатных изданий, но не позволяющие осуществлять ни узко специальных, ни комплексных исследований летописных списков.

Понятно, что переход от веб-библиотек, содержащих сканированные изображения страниц, машиночитаемые копии отдельных текстов, рукописей и их коллекций, к имеющим глубокую разметку многофункциональным корпусам не может вызывать сомнений сегодня в связи с необходимостью создания основы для анализа исторических документов современными, в частности лингвотекстологическим и корпусными методами.

Одним из наиболее перспективных направлений в области подготовки лингвистических ресурсов на основе средневековых письменных источников, безусловно, является создание параллельных корпусов, содержащих различные варианты одного текста-произведения. Традиционно подобные корпусы создаются для текстов, переведенных на другие языки. Для историко-лингвистических исследований не менее значимым является создание и использование параллельного корпуса на основе разновременных списков одного произведения, содержащих важные для исторического, лингвистического, литературоведческого анализа расхождения. Именно такие корпусы могут дать широкому кругу интересующихся конкретным, значимым для политической, социальной, культурной, художественной, письменной истории народа произведением, существующим в нескольких списках, различающихся составом, структурой и языком, полные

сведения о его вариантах, а ученым – иметь материал для изучения истории создания текста, его редактирования и бытования.

Известно, что корпусные методы позволяют ставить и решать задачи, требующие достаточной частотности явлений, например, в графике, орфографии, пунктуации, фонетике, морфологии. Анализ этих явлений на базе параллельного корпуса разновременных списков одного текста дает необходимую информацию, позволяющую представить в первую очередь в количественных данных вариативность альтернативных языковых единиц во времени, а тем самым обнаружить трудно уловимые с помощью традиционных методов существенные изменения языковой системы.

В настоящее время на портале «Манускрипт» уже существуют коллекция русских летописей и первая версия параллельного корпуса, содержащие три полных списка – Лаврентьевский, Ипатьевский, Радзивиловский (http://manuscripts.ru/mns/portal.main?p1=23). Созданная полнотекстовая база данных рукописей и возможности информационно-аналитической системы «Манускрипт» (система «Манускрипт») (портал проекта: http://manuscripts.ru/) позволяют адекватно представить все графико-орфографические особенности древнерусских рукописей, а также осуществить с помощью многотекстового запросного модуля выборку необходимого для лингвистического анализа лексического, морфологического, словообразовательного материала и визуализировать его в виде прямого, количественного и обратного указателей не только по одной, но и по нескольким или всем рукописям коллекции (о технологиях и возможностях системы "Манускрипт" см., например, [Baranov 2007; Баранов 2008а; Баранов 2008б; Баранов 2012]).

Вторая версия параллельного корпуса будет включать дополнительно к имеющимся еще две транскрипции – Синодальный и Комиссионный списки Новгородской первой летописи, несколько видов выравнивания – по погодным записям, по типам изложения, по писцам и по времени создания фрагмента, а также усовершенствованный веб-инструментарий для создания и демонстрации выборок (о модуле параллельных корпусов системы "Манускрипт" см. [Баранов–Гнутиков 2008; Баранов–Дубовцев 2010]).

Аналитическая часть работ по созданию корпуса связана с текстологическим анализом текстов для установления границ соответствующих друг другу фрагментов. Помимо параллельного корпуса на основе погодных записей, нахождение границ которых в целом не составляет сложности в связи с наличием в списках указания на год записи, корпус создается на основе параллелей жанрово-

содержательных фрагментов текстов, что требует выделения и установления типа изложения каждой такой части, выявления в тексте их границ на основе дифференцирующих типы изложения параметров и нахождения соответствующих друг другу фрагментов в разных списках. Так как в списках часто отсутствуют формальные границы таких частей, эта работа требует применения текстологических и лингвистических методик анализа (теоретические основы выделения типов изложения изложены в [Килина–Зайнуллина, 2010]).

Для осуществления лингвотекстологических исследований корпус размечается также по писцам и времени создания тех или иных фрагментов рукописей, для чего используются уже имеющиеся в науке сведения о количестве писцов каждого списка и о наличии записей, приписок, добавлений в них. В то же время подготовка этой разметки показала, что в работах не всегда указываются точные начало и конец такого рода частей. Поэтому требуется палеографический анализ для установления границ между фрагментами с точностью до знака.

Прикладная часть работ заключается в сверке, редактировании электронных копий и в их разметке. Работы выполняются с помощью специализированного редактора Olded и модуля фрагментирования системы “Манускрипт” (http://manuscripts.ru/mns/cred.analyzer?koll=62133570&f_type=14001). Редактор обеспечивает ввод и редактирование транскрипций рукописей непосредственно в базе данных системы. Кодово-шрифтовая система, включающая все необходимые буквы и их варианты старославянского алфавита, диакритические знаки, титла, пунктуационные и другие символы, позволяет создавать копии, максимально близкие к оригиналу [Баранов–Романенко 2009]. Редактор позволяет вручную размечать тексты в соответствии с имеющимися в системе типами фрагментов, в частности создавать в транскрипции единицы *Погодная запись, Тип изложения, Писец, Время создания* и устанавливать связь фрагмента с соответствующим ему фрагментом другой транскрипции, или присваивая им идентичные значения ключевых свойств, или используя перечень (словарь) фрагментов данного типа (более подробно см. [Редактор 2009]). Автоматизированная разметка (фрагментирование) списков одного текста может быть осуществлена с помощью модуля фрагментирования, который автоматически ищет в целевых транскрипциях фрагменты, соответствующие фрагменту основной рукописи, позволяет указать точные границы найденных частей и сохранить в базе данных новый фрагмент и связь между целевым и найденными фрагментами (более подробно см. [Баранов 2011а]).

Демонстрация параллельного корпуса в Интернете осуществляется с помощью процедур модуля параллельных корпусов системы “Манускрипт”. Особенностью модуля является возможность выбора визуализируемой части текста на основе указания основного списка, диапазона его листов или конкретных фрагментов, возможность выбора пользователем формы представления соответствующих друг другу фрагментов – только заголовки фрагментов, инципит и эксплицит фрагмента или полный его текст. В настоящее время ведутся работы по увеличению количества параметров запроса за счет добавления поиска по нескольким маскам лингвистических единиц, по расстоянию между ними, по грамматическим значениям словоформ, лемматизация которых будет осуществлена с помощью морфологического анализатора системы (см. [Баранов и др. 2007; Baranov 2008]), планируется совершенствование дизайна запросных веб-форм и веб-форм визуализации выборок (о модуле параллельных корпусов и критических изданий см., например, [Баранов–Гнутиков 2008; Баранов–Дубовцев 2010; Баранов 2012]).

Полная разметка по погодным записям, по типам изложения, по писцам и времени создания фрагментов существенно расширяет возможности использования полнотекстовой базы данных летописей. В частности, многотекстовый модуль предоставляет возможность создавать перечни лингвистических единиц, входящих в тот или иной тип фрагмента, что позволяет ставить и решать задачи сопоставительного анализа списков одного произведения лингвотекстологическим методом (об опыте использования модулей для историко-лингвистических исследований см., например, [Аникина 2010; Аникина 2012а; Аникина 2012б; Баранов 2010; Baranov 2010; Баранов 2011б]), прототип модуля статистики (http://manuscripts.ru/mns/cred.stat) дает возможность наблюдать за изменением частоты использования лингвистических единиц в пределах заданного диапазона фрагментов того или иного типа (о возможностях модуля см., например, [Баранов 2012]).

Таким образом, создаваемый параллельный корпус древнейших русских летописей позволяет ставить и решать широкий круг фундаментальных задач исторической русистики в области грамматики, лексики, словообразования корпусными и лингвотекстолическим методами, а также использовать материалы корпуса в образовательном процессе в качестве учебного, иллюстративного и/или исследовательского материала в исторических и историко-лингвистических учебных курсах.

Благодарности

Работа выполняется при финансовой поддержке Министерства образования и науки РФ в рамках государственного задания на выполнение работ ФГБОУ ВПО "Ижевский государственный технический университет" (проект № 8.1613.2011 “Средневековый славянский текст как объект текстологического, лингвистического и структурного моделирования: обеспечение миграции полнотекстовых машиночитаемых исторических документов”).

Список источников параллельного корпуса

Лаврентьевская летопись, РНБ, F.п.IV.2, 1377 г., 146 л.

Ипатьевская летопись, БАН, 16.4.4, перв. пол. (сер.?) XV в., 307 л.

Радзивиловская летопись, БАН, 34.5.30, XV в., 245 л.

Новгородская первая летопись по Синодальному списку, ГИМ, Син. 786, XIII в., 169 л.

Новгородская первая летопись по Комиссионному списку, ИРИ РАН (СПб), Арх., 240, XV в., 320 л.

Литература

Аникина 2010 – Аникина Р.А. Именные формы на -ьн(о) в Повести временных лет (корпусный и лингвотекстологический подход) // Межд. молодежный научный форум "Ломоносов–2010": Материалы XVII Межд. науч. конф. студентов, аспирантов и молодых ученых "Ломоносов". Секция "Филология" (Москва, 12–15 апреля 2009 года). М.: МАКС Пресс, 2010. 1 электрон. опт. диск (CD–R).

Аникина 2012а – Аникина Р.А. Разночтения в употреблении именных форм на -о как показатель формирования приглагольного определителя в древнерусском языке (на материале корпуса русских летописей XIV–XV вв. и корпуса славянских евангелий XI–XIII вв.) // Вестник Орловского государственного университета. Орел, 2012. № 4(24). (Серия "Новые гуманитарные исследования").

Аникина 2012б – Аникина Р.А. К вопросу о грамматической характеристике форм на –ьн(о) в древнерусском языке (на материале летописных и евангельских

текстов) // Русский язык: функционирование и развитие (к 85-летию со дня рождения заслуженного деятеля науки Российской Федерации профессора Виталия Михайловича Маркова): материалы Межд. науч. конф. (Казань, 18–21 апреля 2012 г.). Казань: Казан. ун-т, 2012. Т. 1. С. 211–220.

Baranov 2007 – Victor A. Baranov. The ideology and technology of creating online full-text digital collections of ancient and medieval slavonic manuscripts // International Conference on Applied Natural Sciences. Trnava (November 7–9, 2007). P. 199–207.

Баранов и др. 2007 – Баранов В. А., Миронов А. Н., Лапин А. Н., Мельникова И. С. и др. Автоматический морфологический анализатор древнерусского языка: лингвистические и технологические решения [Электронный ресурс] // 10-я юбил. межд. конф. «EVA 2007 Москва». Москва, 2007. URL: http://conf.cpic.ru/eva2007/rus/reports/report_1130.html (дата обращения: 03.06.2012).

Баранов 2008а – Баранов В. А. Проект «Манускрипт»: предварительные итоги // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам: материалы междунар. науч. конф. (Казань, 26–30 августа 2008 г.) / отв. ред. В. А. Баранов, В. Д. Соловьев. Казань: Изд-во КГУ, 2008. С. 32–36.

Баранов 2008б – Баранов В. А. Полнотекстовые базы данных как основа для электронных изданий средневековых рукописей в Интернете: требования, реализация, перспективы // Scripta & e-Scripta: The Journal of Interdisciplinary Mediaeval Studies. Vol. 6. Sofia: “Boyan Penev” Publishing Center; Institute of Literature, BAS, 2008. С. 47–64, 422.

Баранов–Гнутиков 2008 – Баранов В. А., Гнутиков Р. М. Электронное критическое издание средневекового текста: постановка задачи, основные требования и инструментальная подготовка // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам: материалы междунар. науч. конф. (Казань, 26–30 августа 2008 г.) / отв. ред. В. А. Баранов, В. Д. Соловьев. Казань: Изд-во КГУ, 2008. С 37–44.

Baranov 2008 – Victor A. Baranov, Aleksey N. Mironov, Aleksey N. Lapin, Irina S. Melnikova [et al.] Development of the Processing and Visualization Technologies for the Linguistic Information in the Manuscript System: Lemmatization // JADT 2008: actes des 9es Journées internationales d’Analyse statistique des Données Textuelles, Lyon, 12–14 mars 2008: proceedings of 9th International Conference on

Textual Data statistical Analysis. Lyon (March 12–14, 2008). / Scientific editors: Serge Heiden, Bénédicte Pincemin. Lyon: Presses Universitaires de Lyon (PUL). Vol. 2. P. 137–145.

Баранов–Романенко 2009 – Баранов В. А., Романенко В. А. Опыт разработки, создания и использования кирилловского алфавита для полнотекстовых баз данных и интернет-изданий древнерусских рукописей XI–XIV веков // Стандардизација старословенског ћириличког писма и његова регистрација у уникоду: Зборник радова са међународног научног скупа одржаног од 15. до 17. октобра 2007. године / Уредници Гордана Јовановић, Јасмина Грковић-Мејџор, Зоран Костић, Виктор Савић. – Београд: Српска академија наука и уметности, 2009. С. 49–62. (Научни скупови: књига CXXV. Одељење језика и књижевности. Књига 20).

Баранов 2010 – Баранов В.А. Корпус средневековых славянских письменных памятников и лингвотекстологические исследования в области исторической морфологии русского языка // Информационные технологии и письменное наследие: материалы междунар. науч. конф. (Уфа, 28–31 октября 2010 г.) / отв. ред. В. А. Баранов. Уфа; Ижевск: Вагант, 2010. С. 21–27.

Baranov 2010 – Victor A. Baranov. Machine-Readable Linguistic Internet Resources as a Basis for Historical-Philological Studies // Journal of Applied Mathematics, Statistics and Informatics. Volume 6, Number 2, December 2010. Trnava: The University of SS. Cyril and Methodius, Faculty of Natural Sciences, 2010. Pp. 63–89.

Баранов–Дубовцев 2010 – Баранов В.А., Дубовцев С.В. Электронное критическое издание средневекового славянского текста: модель данных и визуализация лингвистических единиц // Интеллектуальные системы в производстве. 2010. № 1. С. 280–287.

Баранов 2011а – Баранов В.А. Полнотекстовая коллекция славянских Евангелий проекта "Манускрипт" и специализированные инструменты разметки: модуль фрагментирования // Вестник Пермского университета. Серия "История". Вып. 2 (16). 2011. С. 40–47.

Баранов 2011б – Баранов В.А. Software Tools and User Interfaces designed for Historical-Linguistic Purposes of Project “Manuscript” // Информационный бюллетень Ассоциации «История и компьютер». №37. Труды междунар. конф. «Компьютерные технологии и математические методы в исторических исследованиях» (Петрозаводск, 11–16 июля 2011 г.). Петрозаводск: 2011. С. 7–14.

Баранов 2012 – Баранов В. А. Лингвистические, методические и технологические вопросы создания и использования корпуса средневековых славянских текстов // Русистика: язык, культура, перевод: сб. докладов юбилейной междунар. науч. конф. (София, 23–25 ноября 2011 г.). София: Изток–Запад, 2012. С. 404–414.

Килина–Зайнуллина 2010 – Килина Л.Ф., Зайнуллина С.Р. Тип изложения как основание для фрагментирования летописного текста // Информационные технологии и письменное наследие: материалы междунар. науч. конф. (Уфа, 28–31 октября 2010 г.) / отв. ред. В.А. Баранов. Уфа ; Ижевск: Вагант, 2010. С. 104–108.

Редактор 2009 – Редактор OldEd: Руководство пользователя / Р. М. Гнутиков, В. А. Баранов. Изд. 2-е, перераб. и доп. Ижевск, 2009. 121 с.