Проект «Манускрипт»: предварительные итоги

В. А. Баранов

Ижевский государственный технический университет, Удмуртский государственный университет, Ижевск, Россия

1. С момента появления в Интернете первого электронного издания «Путятина минея», созданного группой лингвистов и программистов Удмуртского университета (URL: http://manuscripts.ru/mns/portal.main?p1=19&p_lid=1), прошло более шести лет. Сегодня портал «Манускрипт: славянское письменное наследие» (URL: http://manuscripts.ru/) содержит несколько коллекций древних славянских письменных памятников.

За время работы коллектива над электронными средневековыми изданиями существенно переработаны специализированные модули ввода, редактирования и обработки текстов, значительно изменились технологии подготовки текстов и справочных материалов к публикации, созданы новые web-сервисы представления данных пользователю, но первоначальная направленность проекта на создание лингвистических ресурсов, позволяющих получать информацию для решения различных задач в области исторического языкознания и смежных гуманитарных дисциплин, осталась без изменений. Прежними остаются требования к структурированию текстовых данных, к результатам их представления в Интернете и технологические решения, позволяющие их достичь:

- хранение текста в базе данных, минимальной единицей которой является символ,
- создание любых формально представленных в рукописях / текстах единиц как совокупности подчиненных объектов,
- наличие у объектов базы данных свойств и значений, соответствующих их текстовым и лингвистическим аналогам,
- создание web-модулей, обеспечивающих распределенную и удаленную работу над электронными изданиями,

- создание web-сервисов, позволяющих конечному пользователю формировать запрос на основе нескольких параметров и получать результат в необходимом ему виде.

Эти подходы позволяют при работе с коллекциями в Интернете обеспечить:

- получение близкого к оригиналу транскрипционного представления текста,

- формирование запроса на основе мета- и аналитических свойств и значений рукописей, текстов и их фрагментов, а также на основе значений лингвистических единиц,

- вывод текстов / рукописей в виде различного вида перечней запрашиваемых единиц и некоторые другие возможности.

2.1. В то же время система «Манускрипт» сегодня представляет собой уже принципиально иной проект по сравнению с тем, который позволил осуществить первое электронное издание. Назовем основные особенности применяемых и разрабатываемых в настоящее время технологий:

- предоставление автору электронного издания инструментов, позволяющих в удаленном режиме подготовить текст к изданию и опубликовать его на портале «Манускрипт: славянское письменное наследие»,

- предоставление автору публикации дополнительных лингвистических ресурсов (морфологических словарей), позволяющих осуществить морфологический разбор и создать словоуказатели,

- предоставление автору инструментов для мета- и аналитического описания текстов и рукописей,

- предоставление конечному пользователю многофункциональных web-модулей, позволяющих осуществить поиск в коллекциях портала и отобрать необходимые для анализа тексты и/или рукописи,

- наличие возможностей сформировать конечное представление выборки по нескольким текстам, рукописям или их фрагментам в виде сравнительных перечней лингвистических и/или структурных единиц и обеспечение переходов от единиц перечней к контекстам,

- возможность поиска текстовых прецедентов и автоматического приведения их к лемме, а также некоторые другие.

2.2. Все перечисленные функции обеспечиваются технологической цепочкой специализированных модулей, среди которых необходимо выделить:

- редактор OldEd, с помощью которого осуществляется набор и редактирование текста, ввод значений единиц и установление связей между ними, лемматизация словоформ и другие операции,

- web-модуль ввода и редактирования свойств и значений единиц базы данных, который, в частности, обеспечивает создание и правку справочников, используемых при разборе текстов и их единиц,

- web-модули грамматических словарей и морфологического анализа, с помощью которых осуществляется лемматизация,

- web-модуль поиска текстов, рукописей и их фрагментов, обеспечивающий формирование запроса на основе мета- и аналитических характеристик,

- web-форма запроса и представления результатов выборки по нескольким текстам / рукописям / фрагментам в виде сравнительных перечней их единиц;

- web-модуль запросов и выборок, с помощью которого может быть осуществлена выборка данных с учетом всех параметров объектов базы данных, произведены различные операции над выборками, а также сформирован и выведен на печать в необходимом пользователю виде результат.

3. Целям и задачам проекта «Манускрипт», лингвистическим и технологическим решениям, достигнутым прикладным и теоретическим результатам посвящено несколько публикаций [Baranov 2004, Baranov 2006a, Baranov 2006b, Baranov 2006c, Gnutikov 2007, Baranov 2007, Баранов 2007] (см. также статьи членов коллектива в [Современные информационные технологии и письменное наследие 2006]), поэтому остановимся только на некоторых важных модулях системы, обеспечивающих перечисленные выше возможности.

3.1. Существенно расширены функциональные возможности специализированного редактора OldEd в направлении создания критического электронного издания – электронной коллекции, содержащей рукописи одного текста. Доработка редактора позволяет в настоящее время создать архетип (антиграф, протограф) текста, лингвистические и структурные единицы которого могут быть связаны в базе данных с соответствующими единицами в конкретных рукописях.

Имевшиеся в редакторе функции по установлению связей между единицами текстов / рукописей / фрагментов и соответствующими единицами словарей и справочников пополнились возможностями автоматизированной лемматизации словоформ с помощью грамматического словаря древнерусского языка.

Из других изменений следует отметить появление возможностей работы с выборками, сформированными в модуле запросов и выборок, навигации по ним и редактирования их единиц, а также средств объединения текстов в коллекции, что позволяет самим авторам изданий формировать и публиковать на портале коллекции.

3.2. Для поиска текстовых прецедентов в коллекциях и для вывода результатов запроса в виде перечня начальных форм создано несколько версий морфологического анализатора древнерусского языка (совместно с сотрудниками Института русского языка им. В. В. Виноградова РАН, руководитель работы – А. А. Пичхадзе) (версия 4 доступна по адресу http://manuscripts.ru/mns/lmtz.search_form_ex). Модель электронного грамматического словаря и правила, применяемые при лемматизации, позволяют осуществлять приведение к начальной форме соответствующих маске текстовых прецедентов вне зависимости от степени их графико-орфографического отличия от единиц словаря. Различные виды представления результата поиска и лемматизации позволяют получить сведения о наличии / отсутствии искомых единиц в текстовых подкорпусах, о их количестве и перейти к соответствующим контекстам.

3.3. В единый модуль объединены web-модуль поиска текстов, рукописей и их фрагментов, обеспечивающий формирование запроса на основе мета- и

аналитических характеристик, и web-форма запроса и представления результатов на основе полученных текстовых единиц в виде текстов и сравнительных перечней их единиц.

Первый модуль реализован в двух вариантах: в виде запросных форм для простого (URL: http://manuscripts.ru/mns/srch.simple) и расширенного (URL: http://manuscripts.ru/mns/srch.complex?p_lang=RU) поиска.

Простой поиск осуществляется с помощью маски искомого значения метахарактеристик текстов и рукописей, а также аналитических описаний их фрагментов. Полученный результат может быть уточнен повторным запросом.

Расширенный поиск дает возможность сформировать запрос на основе нескольких параметров – различных свойств и значений рукописей / изданий, текстов / произведений и фрагментов (возможно использование логических *и*, *или*).

Сохранение пользовательских выборок позволяет использовать их неоднократно.

Во второй форме на основе отобранных рукописей, текстов или фрагментов может быть сформировано необходимое пользователю представление – в виде текста(ов) или упорядоченных перечней их объектов. Идентификация конкретной единицы осуществляется посредством указания адреса, который позволяет перейти к соответствующему месту рукописи.

4. Сегодня на портале представлено несколько коллекций древних славянских письменных памятников, основной из которых следует считать коллекцию рукописей, созданных на Руси в XI веке (URL: http://manuscripts.ru/mns/portal.main?p1=27). Это транскрипции двадцати рукописей различного объема: это и списки полных произведений от ста до трехсот и более листов, и отрывки в несколько листов. Общий объем текстовых прецедентов в этой коллекции – более 500 000 словоупотреблений. Набор, сверка и редактирование коллекции были осуществлены в 1991–2007 годах сотрудниками, аспирантами и студентами УдГУ и ИжГТУ. В настоящее время проверка транскрипций по микрофильмам продолжается совместно с преподавателями Софийского университета (руководитель группы – Румяна Павлова).

Другие коллекции содержат древнейшие русские списки Евангелий, служебных миней, триодей, три списка Повести временных лет, несколько русских переводов и несколько святоотеческих произведений.

Созданием коллекций средневековых произведений не ограничиваются интересы лингвистов коллектива. В настоящее время возобновлены и совместно с коллегами из Казанского госуниверситета (руководитель – В. Д. Соловьев) ведутся работы по созданию коллекции М. В. Ломоносова (URL: http://manuscripts.ru/mns/portal.main?p1=31).

Благодарности

Теоретические и прикладные работы по созданию системы «Манускрипт» выполняются в рамках проектов № 07-04-00369а (исследование майских служебных миней), № 07-04-12147в (создание коллекции М. В. Ломоносова), № 07-04-12140в (портал «Манускрипт: славянское письменное наследие»), поддержанных Российским гуманитарным научным фондом, а также в рамках тематического плана Ижевского государственного технического университета (специализированные модули системы).

Литература

Baranov 2004 – *Baranov, Victor*. Old Slavic Manuscript Heritage: Electronic Publications and Full-Text Databases / Victor Baranov, Andrey Votintsev, Roman Gnutikov, Aleksey Mironov, Sergey Oshchepkov, Vitaliy Romanenko // EVA 2004 London (Electronic Imaging, the Visual Arts Conference & Beyond) : Conference Proceedings / University College London. Institute of Archaeology ; principal Editor James Hemsley. – London, 2004. – P. 11.1-11.8.

Современные информационные технологии и письменное наследие 2006 – *Современные информационные технологии и письменное наследие*: от древних рукописей к электронным текстам : материалы междунар. науч. конф., Ижевск, 13–17 июля 2006 г. / Отв. ред. В. А. Баранов. — Ижевск : Изд-во ИжГТУ, 2006. — 193 с.

Baranov 2006a – *Baranov, Victor.* An Editor of Ancient Texts as Part of the System «Manuscript» [Текст ; Электронный ресурс] / Baranov, Victor Arkadievich; Votintsev, Andrey Anatolievich; Gnutikov, Roman Michailovich // ELPUB2006. Digital Spectrum: Integrating Technology and Culture : Proceedings of the 10th International Conference on Electronic Publishing held in Bansko, Bulgaria, 14-16 June 2006 / Edited by Bob Martens, Milena Dobreva. – 2006. – P. 375–376. – Режим доступа : http://elpub.scix.net/cgi-bin/works/Show?267_elpub2006, свободный. – Загл. с титул. страницы.

Baranov 2006b – *Baranov, Victor*. Information-Analytical System “Manuscript”: technologies and tools of creation of electronic collections of ancient and medieval documents [Электронный ресурс] / Victor Baranov // Dagstuhl Seminar Proceedings 06491: Digital Historical Corpora - Architecture, Annotation, and Retrieval / L. Burnard, M. Dobreva, N. Fuhr, A. Lüdeling; Dagstuhl Seminar 06491, 03.12. – 08.12.2006; Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany. – Режим доступа : http://drops.dagstuhl.de/portals/index.php?semnr=06491, свободный. – Загл. с титул. страницы.

Baranov 2006c – *Baranov, Victor*. The Computer Model of the Slavonic Text: Syntactic Units, relationships and Properties in the Database of the Manuscript System / Victor Baranov // Computer Applications in Slavic Studies : Proceedings of Azbuky.Net : International Conference and Workshop, Sofia, Bulgaria, October 24–27, 2005. – Sofia : “Boyan Penev” Publishing Center; Institute of Literature, Bulgarian Academy of Science, 2006. – P. 61–68.

Gnutikov 2007 – *Gnutikov, Roman*. Up-to-date means of access to full-text databases / Roman Gnutikov, Victor Baranov // Digital Humanities 2007 : Conference Abstracts / The 19th Joint International Conference of the Association for Computers and the Humanities, and the Association for Literary and Linguistic Computing, at the University of Illinois, Urbana-Champaign, USA, June 4 – June 8, 2007. – Urbana-Champaign : Graduate School of Library and Information Science; University of Illinois, 2007. – P. 74–76.

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

Баранов 2007 – Баранов, В. А. Автоматический морфологический анализатор древнерусского языка: лингвистические и технологические решения [Электронный ресурс] / В. А. Баранов, А. Н. Миронов, А. Н. Лапин, И. С. Мельникова, А. А. Соколова, Е. А. Корепанова // «EVA 2007 Москва» : 10-я юбилейная междунар. конф. – Москва, 2007. – Режим доступа : http://conf.cpic.ru/eva2007/rus/reports/report_1130.html, свободный. – Загл. с титул. страницы.

The Manuscript projects: preliminary results

Victor A. Baranov

Izhevsk State Technical University, Udmurtia State University Izhevsk, Russia

This paper describes the Manuscript informational-analytical system, which is designed for storage, processing, and online publication of medieval Slavic manuscripts and their reference apparatus. Special attention is given to new modules enabling lemmatization and to new methods for presenting units of the corpus – comparative lists that can be used to obtain data necessary for linguistic and textological investigations.