МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. М. Т. КАЛАШНИКОВА

Информационные технологии и письменное наследие

El'Manuscript–2012

Материалы IV международной научной конференции
Петрозаводск, 3–8 сентября 2012 года

Петрозаводск, Ижевск
2012

УДК 004.9
ББК 81.11+81.2-0
И741

Изданы при поддержке гранта РФФИ (проект № 12–06–06061–г),
гранта РГНФ (проект № 12–04–14154–г) и в рамках реализации комплекса
мероприятий Программы стратегического развития ПетрГУ на 2012-2016 г.

Ответственные редакторы:
В. А. Баранов, д-р филол. наук, проф.
А. Г. Варфоломеев, канд. физ.-мат. наук, доц.

И741 **Информационные технологии и письменное наследие** [Текст] : материалы IV междунар. науч. конф. (Петрозаводск, 3–8 сентября 2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. — Петрозаводск ; Ижевск, 2012. — 328 с.

ISBN 978–5–8021–1402–5

Сборник содержит материалы конференции, посвященной современным электронным средствам хранения, описания, обработки, исследования и публикации памятников письменности и исторических источников.

УДК 004.9
ББК 81.11+81.2-0

ISBN 978–5–8021–1402–5

КОРПУС ЦЕРКОВНОСЛАВЯНСКИХ ТЕКСТОВ В СОСТАВЕ НКРЯ, ПЕРВАЯ ВЕРСИЯ: ПРОБЛЕМЫ И РЕШЕНИЯ

А. Е. Поляков1, Е. Р. Добрушина2, Т. Ю. Иванова-Алленова2
1НПБ им. К.Д. Ушинского, Москва;
2Православный Свято-Тихоновский Гуманитарный Университет, Москва

The paper discusses the issues of spelling, grammatical tagging and metatext markup for a pilot version of the Church Slavonic Corpus (started in 2011) and the prospects of its development

1. Корпус церковнославянских текстов[1]. За 2011 год группой иссле-дователей[2] создан и размещен в Интернете (пока идет отладка — на адре-се http://ruscorpora.ru/beta/search-orthlib.html) пилотный вариант представи-тельного по объему корпуса специальным образом подготовленных церков-нославянских текстов. Тексты снабжены специальной разметкой (метатек-стовой и грамматической), их общий объем — 1250 текстов, около 4,6 мил-лиона словоупотреблений, включающих около 150 тыс. различных слово-форм. Данный корпус является новой частью проекта "Национальный кор-пус русского языка" (http://ruscorpora.ru/).

2. Тексты корпуса и метатекстовая разметка. Основу корпуса цер-ковнославянских текстов в составе НКРЯ составляют современные богослу-жебные тексты (XIX–XX вв.) (60%). Кроме того, в корпусе представлены другие периоды (XVII–XVIII вв.) и жанры церковнославянских текстов: пи-сание, святоотеческие, правовые, научные, и др.

Жанрово-тематическая классификация текстов, на основе которой поль-зователь может создать пользовательского подкорпус, адресована, в первую очередь, массовому пользователю, а не узкому специалисту, и строится ис-ходя из принципов практического удобства и понятности. Метатекстовая разметка, основная концепция которой разработана А.Г. Кравецким, в на-стоящее время проводится только на уровне целого текста и состоит из сле-дующих параметров (перечисляются корпусные «ярлыки»):

– «Писание» — это Библия, Служебное евангелие и подборки пари-мий в богослужебных книгах, если они выделены в отдельную рубрику;
– «святоотеческий»;
– «типикон»;
– «служба» — это все богослужебные чины и службы, а также под-борки богослужебных текстов (богородичны, кондаки и т. д.) в составе раз-ных сборников;
– «акафист»;
– «право» — это тексты по церковному праву, например, «Книга Правил святых апостолов, святых соборов, вселенских и поместных, и свя-тых отцов»;
– «научный» — например «Ифика Иерополитика».

В будущем предполагается создать метаразметку на уровне отдельных фрагментов текста (например, кондаки, тропари), которые входят в состав больших текстов.

[1] Авторы выражают благодарность за помощь в создании церковнославянского корпуса Про-грамме фундаментальных исследований Президиума РАН «Корпусная лингвистика». Также авторы благодарят за существенную помощь компанию Яндекс, в компьютерном простран-стве которой существует весь Национальный корпус русского языка.

[2] Помимо авторов данных тезисов в создании церковнославянского корпуса на разных этапах его разработки в качестве создателей концепции, разработчиков и исполнителей технических задач, а также консультантов и вдохновителей работы принимали участие В.А.Плунгян, А.Г.Кравецкий, А.И.Зобнин, А.В.Жирова, А.А.Плетнева, Л.И.Маршева, свящ. Ф.Б.Людоговский, Р.Н.Кривко, И.В.Сегалович, свящ. К.О.Польсков.

Также доступен отбор текстов по периоду создания, точнее, по следующим параметрам:

– «архаичный тип» — например, «Добротолюбие»;

– «гибридный тип» — например, «Алфавит Духовный»;

– «стандартный тип» — это все тексты основных богослужебных книг, за исключением текстов XX века;

– «XX век» — это, в первую очередь, акафисты, например «Всем святым, в земле Российской просиявшим».

К сожалению, нет поиска по конкретной дате создания или издания тек-ста, так как определение таких дат для многих текстов требует кропотливой работы, а иногда и вовсе не представляется возможным, более того, для тех текстов, где даты ясны, невозможно определить, в какой момент — в момент создания или в момент издания — производилась последняя языковая правка.

3. Кодировка текста — проблемы орфографии. В процессе подготов-ки были выработаны принципы и правила представления церковнославян-ских текстов в корпусе, а также решены практические проблемы, касающие-ся разметки и кодировки текста. Каноническая церковнославянская орфография, ориентированная на типографское представление текста, включает много символов, которые фактически не несут смыслоразличительной функции. Для представления корпуса в интернете была выработана более простая орфографическая система, которая сохраняет существенные языковые различия, но не пытается имитировать точный типографский вид текста. В этой орфографии отсутствуют некоторые символы с малой или нулевой различительной способностью — придыхания, различие *оу/у*, *ia/я*, — однако сохраняются особенности, связанные с различением лексического значения — *о/омега*, *и/i/v*, *ф/фита*, *з/s*. Кроме того, были выработаны правила перевода текста из этой орфографической системы в современный вид для удобства поиска.

Словоформы в корпусе представлены так же, как в исходном тексте, а леммы даются в унифицированном написании, в котором не используются избыточные буквы, а титла раскрыты (например, слово *мл^{с}ть* рассматривается как форма к лемме *милость*).

4. Грамматическое описание. Грамматическое описание церковнославянского словоизменения необходимо для работы лексико-грамматического поиска, а в перспективе и для создания автоматического морфологического анализатора церковнославянского языка.

Существующие грамматики и словари описывают некоторую идеальную картину, которая отражает нормализаторские устремления авторов, но часто не соответствует фактическому состоянию языка. Реально в церковнославянских текстах наблюдается множество орфографических и словоизменительных вариантов, обусловленных тем, что тексты создавались в разное время и в разной языковой среде. Поэтому, если исходить из изучения грамматик и словарей, часто нельзя понять, какие формы имеет некоторое слово. В таких случаях единственным достоверным источником является корпус.

Например, считается, что для отличия мн. и дв. числа от омонимичных форм ед. числа используется облеченное ударение (^) и замена *о→w, е→є*. Спрашивается, как будет им. мн. от *высота — высоты^, высwты'* или *высwты^* ? Оказывается, в корпусе встречаются все три формы, причем первая 9 раз, вторая — 17, третья — 1. Из анализа других аналогичных случаев мы можем сделать вывод, что при конкуренции правил замена *о→w* предпочтительнее, чем облеченное ударение.

К настоящему моменту была проделана значительная работа, результатом которой явилась пилотная версия грамматического словаря. Было проанализировано около 150 тыс. словоформ, большинству из них была приписана основная грамматическая информация: лемма, часть речи, грамматические признаки. Вначале была сделана ручная лемматизация для ряда словоформ, а затем на основе построенной модели словоизменения были проанализированы остальные словоформы. Часть словоформ была квалифицирована как ошибки (опечатки). Для некоторых редких форм лемму определить не удалось или она была определена гипотетически. Поиск по переменным характеристикам, таким как падеж или род (для прилагательного), работает, но без снятия грамматической омонимии.

Грамматические таблицы, на которые опирается автоматический анализатор, строятся на основе анализа корпуса текстов и существующих грамматик. Поскольку грамматики дают неполную и противоречивую картину словоизменения, то только корпус является окончательным критерием истины.

В отличие от традиционных грамматик, парадигмы не были заданы априорно, а выведены эмпирически на основе анализа множества словоформ, имеющих однотипное соотношение между грамматическими формами. Таким образом, номенклатура парадигм получается значительно более детальной, чем традиционная, и подчас существенно от таковой отличается. Например, традиционное первое склонение (*рабъ*) на самом деле распадается на 14 подтипов в зависимости от конечного согласного (парный твердый – мягкий, велярный, шипящий, йот), наличия беглого гласного и других особенностей (*іерей – іерея,іереомъ, агарянинъ – агаряне)*.

5. Перспективы церковнославянского корпуса. Чтобы корпус мог в полной мере удовлетворять потребностям современной науки и образования в области исследования и преподавания церковнославянского языка, требуется значительная доработка и развитие. Хотелось бы в будущем сделать следующее:

– Снабдить корпус справочными материалами об особенностях корпуса и составе текстов, облегчающими пользователям работу с корпусом.

– Снабдить корпус кратким словарем, поясняющим термины, использованные в разметке текстов.

– Выверить грамматические характеристики словоформ, приписанные компьютерными методами, и устранить ошибки.

– Приписать леммы и грамматические характеристики словоформам, оставшимся неразобранными в результате программного анализа.

– Произвести путем анализа необработанных слов и их значений в текстах поиск ошибок, возникших при наборе текстов, и выправить эти ошибки.

– Проанализировать значение лемм, определенных разметчиками как имена нарицательные, и приписать тем из них, перевод которых на русский язык не ясен без привлечения специальных знаний, краткое толкование, доступное пользователю вместе с грамматической характеристикой при нажатии на выбранное слово на странице результатов поиска.

– Проанализировать значение лемм, определенных разметчиками как имена собственные, подтвердить или отвергнуть это решение и определить для каждого имени собственного лексико-семантический класс: имя, топо-ним и др. Такая работа позволит расширить представления о лексическом составе церковнославянского языка и употребляющихся в нем имен собст-венных, даст возможность ликвидировать ошибки массовой разметки, а также даст возможность пользователям легче понимать содержание полу-ченного текста.

– Снабдить корпус списком основных лемм и создать для пользователя возможность переходить к поиску интересующей его леммы непосредственно из алфавитного списка, что во многом решит проблемы, связанные с вариативностью орфографии и наличием в запросах букв, требующих специальных шрифтов или использования виртуальной клавиатуры.