МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

Писменото наследство и информационните технологии [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.

ISBN 978–954–9787–25–2

Система электронной грамматической разметки древнерусских и церковнославянских текстов и её использование в веб-ресурсах

Т. А. Архангельский, Е. А. Мишина, А. А. Пичхадзе

Электронная разметка, представление лингвистической информации в Интернете, поисковая система

A System for Digital Morphological Tagging for Old Russian and Church Slavonic Texts and Its Use in Web Resources

Timofey Arkhangelsky, Ekaterina Mishina, Anna Pichkhadze

At the Vinogradov Institute of Russian Language of the Russian Academy of Sciences, a system has been created that provides a graphical interface for manual and semi-automatic tagging (based on precedents) of Old Russian and Church Slavonic texts (both original and translated) and includes a search engine as well as the ability to automatically build lexical and grammatical indices to the text. A web interface (at www.ruslang.ru) for publishing annotated texts online without the help of computer programmers has also been designed.

В Институте русского языка им. В. В. Виноградова (ИРЯ РАН) созданы электронные базы данных по древнерусским и церковнославянским памятникам (древнерусские летописи, берестяные грамоты, переводные и оригинальные церковнославянские тексты) в системе, представляющей собой графический интерфейс для ручной и полуавтоматической (с использованием прецедентных разборов) разметки текстов. Система обеспечивает поиск по размеченному тексту, а также возможность автоматического построения лексико-грамматических указателей (славяно-иноязычный, иноязычно-славянский, обратный словник). Тексты, обрабатываемые в программе, хранятся в файлах формата YAML. Структура данных этого формата похожа на широко используемый формат XML, но при этом файл занимает мало места в памяти и прост для понимания человеком, что позволяет быстро исправлять ошибки в тексте в случае их возникновения. Хранение текстов в этом формате имеет ряд преимуществ: структура данных довольно проста, что снижает вероятность появления ошибок при автоматической обработке; выделяемые в тексте фрагменты, состоящие из нескольких словоформ, могут быть привязаны к этим словоформам, что исключает рассинхронизацию между разными слоями представления текста.

Предусмотрена автоматическая разметка текстов по прецедентным разборам с последующим редактированием, что позволяет существенно сократить время обработки новых текстов. Система, с одной стороны, имеет жесткую связь между отдельными элементами (в первую очередь между текстом и грамматической

аннотацией), а с другой стороны, является достаточно гибкой и предоставляет пользователю свободно конструировать элементы, необходимые для адекватной разметки разного рода текстов. Новый подход, реализованный при разработке системы, состоит в том, что словоформа может иметь любое количество альтернативных грамматических разборов. Прецеденты использования альтернативных разборов при разметке древних текстов нам не известны. Между тем альтернативный разбор часто бывает необходим при работе с древними памятниками, элементы текста которых бывают искажены, содержат лакуны и/или могут неоднозначно интерпретироваться.

Ряд специфических особенностей разрабатываемой системы связан с тем, что она предназначена для работы не только с оригинальными, но и переводными текстами. Она обеспечивает возможность присваивать фрагментам славянского текста переводные эквиваленты. Альтернативный разбор применительно к переводным памятникам позволяет соотнести слова в славянском переводе с вариантами слова в греческих источниках (например, разных списках данного памятника), если эти варианты синонимичны и имеют равные шансы служить эквивалентами славянской лексемы.

В разрабатываемой системе осуществлен новый подход к организации фрагментов. Под фрагментами понимаются комплексы словоформ, объединяемые по разным признакам. Сюда относятся аналитические формы (перфект, плюсквамперфект, сослагательное наклонение, сложные будущие времена и др.), словосочетания, а также — для переводных памятников — совокупность словоформ, соответствующих единице (которая также может представлять собой фрагмент) оригинального текста. Фрагменты привязаны к словоформам, которые в них входят, с тем, чтобы при изменении одной из этих словоформ автоматически изменялся и сам фрагмент. Система позволяет в ручном режиме уточнить характер связи между словоформой и фрагментом, в который она входит, если таким образом достигается более точный и адекватный разбор. Словоформы можно находить по грамматическим характеристикам фрагментов, в которые они входят. Одним из преимуществ является возможность объединять несколько словоформ во фрагмент и присваивать некоторые характеристики не отдельным словоформам, а всему фрагменту в целом. Это позволяет находить, например, причастные формы на *-лъ* в составе перфектов как по запросу “Форма = причастие” (свойство словоформы), так и по запросу “Время = перфект” (свойство всего фрагмента), а также устраняет возможность неполной разметки таких времён и наклонений.

Принципиальной особенностью этой системы является наличие двух словарей лемм как для славянского текста, так и для иноязычного оригинала (в случае переводных памятников). Один словарь привязан к аннотируемому тексту и конструируется каждый раз заново при каждом новом запуске системы. Этот словарь отражает лексикон только данного текста. Второй словарь представляет собой сводный словарь уже размеченных текстов и используется для разметки новых текстов. Предусмотрены механизмы сравнения словарей к отдельным памятни-

кам (словарей первого типа) и редактирования сводного словаря, поскольку он составляется из словарей разных текстов, а представление лемм в словарях к разным текстам может отличаться. Например, омонимы в разных словарях могут получить отличающиеся толкования, которые должны быть унифицированы в сводном словаре.

Генерирование словоуказателей отвечает двум задачам: исследовательскому поиску и изданию текста. Можно генерировать словоуказатели на любой стадии работы над текстом. Все изменения, внесенные при редактировании, автоматически отражаются в указателях. Учитывая специфику древних текстов, предусмотрена возможность ранжировать вручную фонетические и орфографические варианты одной и той же грамматической формы. Накопленная информация о порядке следования вариантов может быть использована при обработке новых текстов; по мере увеличения материала сортировка становится в значительной степени автоматической.

Для публикации полных размеченных древнерусских текстов в Интернете на сайте Института по адресу http://bases.ruslang.ru создан веб-интерфейс с системой хранения данных, возможностью последующего редактирования и осуществления поиска. Созданы модуль экспорта в формат Системы веб-разметки, а также система автоматизированной загрузки размеченных текстов в Интернет. Данная система позволяет загружать размеченные тексты в Систему веб-разметки и редактировать информацию о загруженных текстах (удалять опубликованные тексты; изменять разметку; синхронизировать опубликованные на сайте тексты с оффлайновыми версиями) без привлечения программиста, что является принципиальным новшеством. Размеченные тексты также интегрированы в систему Национального корпуса русского языка (http://www.ruscorpora.ru), в рамках которого возможен поиск по древнерусскому подкорпусу.