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

# МОРФО-СИНТАКСИЧЕСКАЯ РАЗМЕТКА ТЕКСТОВ КОРПУСА СКАТ

*И. В. Азарова, Е. Л. Алексеева*
*Санкт-Петербургский государственный университет, Санкт-Петербург*

SCAT, a digital corpus of Old Russian hagiographic texts, maintained by the Department of Mathematical Linguistics of Saint-Petersburg State University, contains texts published in PDF and XML formats. Work is under way to provide all texts with morphosyntactic tagging as recommended by TEI guidelines (P5).

На кафедре математической лингвистики Санкт-Петербургского государственного университета создан и постоянно пополняется корпус агиографических текстов (СКАТ)1, в котором представлены тексты древнерусских житий по рукописям XVI–XVIII вв.

Тексты, представленные в корпусе, прошли через трудоемкую процеду-ру подготовки: предварительного анализа ее графемного состава, деления текста на слова, предполагающего морфо-синтаксический и семантический анализ текста, представления текста рукописи в электронной форме. Все эти этапы вызывают значительное количество проблем, которые решает исследовательский коллектив СКАТ. Результат анализа текста рукописи мы представляем в публикации, которая содержит текст рукописи с подстрочными примечаниями, комментирующими неясные места (они приводятся в тексте в оригинальном виде), что позволяет читателю понять смысл текста. На сай-те СКАТ затем публикуются рукописи в виде pdf-файлов, они доступны для общего пользования. Кроме того, тексты рукописей представлены в xml-формате, который позволяет преобразовать их в другой формат, используемый сторонними пользователями.

Базовые xml-файлы включают воспроизведение графемного состава ру-кописи на том уровне, который коллектив СКАТ счел информативным [Алексеева, 2009]. Все выделенные слова рукописи снабжены числовыми идентификаторами, что позволяет однозначно определить вхождение слова в определенный текст. Помимо полного графемного представления xml-файл содержит представление слова в упрощенной графике, которое используется при поиске в словоуказателе по корпусу [Азарова и Алексеева, 2008].

С 2006 г. тексты рукописей СКАТ сопровождаются наборами морфоло-гических характеристик. Для каждой словоформы текста указывается ее час-теречная принадлежность и приводятся значения всех релевантных грамма-тических категорий. В формате грамматической разметки предусмотрена возможность отражения переходных явлений: через косую черту приводятся ожидаемое значение соответствующей категории (тип склонения, падеж и т.п.) и реально встретившееся в тексте. Например, тип склонения o/u для существительного доуховъ обозначает, что оно относится к типу склонения на *-ŏ, но имеет окончание типа склонения на *-ŭ.

В соответствии с рекомендациями TEI разработан и опробован шаблон представления морфологической аннотации слова в формате XML: используется атрибут ana, в котором указываются ссылки на идентификаторы соответствующих грамматических свойств из библиотеки свойств (например, ana=”#noun #sing #feminine #genitive …”).

Наличие морфологической разметки текстов корпуса позволяет расширить возможности поиска по корпусу: в качестве поискового запроса пользо-ватель будет иметь теперь возможность задать любое сочетание признаков слова, имеющихся в системе.

Морфологическая разметка текстов проводится вручную, в рамках лин-гвистической практики студентов, и затем выверяется квалифицированным специалистом коллектива СКАТ. Мы исходим из того, что использование автоматической разметки при наличии неустойчивой орфографии и поэтому высокой вариативности написания слов не будет давать сколько-нибудь на-дежных результатов аннотации. Таким образом, морфологическая разметка текстов является однозначной.

С 2010 г. мы приступили к разработке формата представления в корпусе синтаксической информации, причем за основу нами был взят перечень син-таксических отношений, используемый в Национальном корпусе русского языка, (который, в свою очередь, был выработан в Лаборатории компьютер-ной лингвистики Института проблем передачи информации РАН). Анализ особенностей старославянского и древнерусского синтаксиса позволил вве-сти соответствующие коррективы: часть отношений была устранена, а целый ряд отношений был добавлен.

**Список литературы**

Азарова и Алексеева, 2008 — Азарова И. В., Алексеева Е. Л. Санкт-Петербургский корпус агиографических текстов (СКАТ): формат XML-представления лингвистической информации и организация поиска данных на сайте // Современные информационные технологии и письменное насле-дие: от древних текстов к электронным библиотекам: Материалы междуна-родной научной конференции. Казань, 2008. С. 3–6.

Алексеева, 2009 — Алексеева Е.Л. Состав графем древнерусских агио-графических текстов // Стандардизација старословенског ћириличког писма и његова регистрација у Уникоду. Зборник радова са међународног научног скупа одржаног од 15. до 17. октобра 2007. године. Београд, 2009. С. 39–48.