

Санкт-Петербургский корпус агиографических текстов (СКАТ): формат XML-представления лингвистической информации и организация поиска данных на сайте

И. В. Азарова, Е. Л. Алексеева

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

На кафедре математической лингвистики Санкт-Петербургского государственного университета в конце 70-х годов XX века начал создаваться корпус агиографических церковнославянских текстов XVI–XVII вв. Работа, в которой учитывались исследования и издания таких текстов, началась с создания картотеки житий святых русской церкви, похвальных слов, сказаний. Был образован фонд фото- и ксерокопий рукописных житий, находящихся в разных хранилищах Петербурга, который пополняется и в настоящее время. В конце 70-х годов начался также ввод житийных текстов в компьютер. В настоящее время в корпусе хранится информация о 52 житиях, общий объем которых составляет более 500 тыс. словоупотреблений.

Было начато изучение грамматики, словообразования и лексики житийных текстов. К середине 90-х годов собранные материалы были обобщены в трех монографиях, которые содержат систематическое описание именного склонения, глагольного спряжения и именного словообразования памятников русской агиографической литературы XVI вв.

С конца 90-х годов на кафедре математической лингвистики издается серия публикаций «Памятники русской агиографической литературы». В каждой книге серии приводится текст одного или нескольких связанных между собой житий. Тексты сопровождаются указателем словоформ, текстологическим очерком, историческими сведениями о святых и основанных ими обителях. Хотя текст жития воспроизводится по одному списку, его выбор отнюдь не случаен. Для всех неясных мест, неизбежно встречающихся в любой рукописи, мы обязательно приводим в аппарате чтения еще двух-трех списков, что позволяет представить текст жития в наиболее полном виде. В издании графика оригинала воспроизводится без какого-либо упрощения: полностью сохраняется используемый писцом набор букв и надстрочных знаков, не раскрываются сокращенные написания под титлом или с

выносными буквами, ошибочные написания комментируются, но не исправляются. Издание отличается от оригинала делением текста на слова и включает указатель словоформ к текстам.

В свет вышло уже 8 выпусков серии, содержащих 13 житий вологодских святых. К изданию готовится 9-ый выпуск.

В 2005 г. для Санкт-Петербургского корпуса агиографических текстов (СКАТ) создан сайт (URL: <http://www.project.phil.pu.ru/scat>), где опубликованные данные размещены для свободного доступа.

На сайте приведена подробная информация о проекте СКАТ, его участниках и публикациях, принципах представления рукописных текстов в корпусе, дается перечень изданий серии и список доступных электронных версий для опубликованных житий.

В настоящее время на сайте доступны два электронных варианта каждого текста. Первый – pdf-представление опубликованного жития, в котором максимально отражены особенности графики, приводятся комментарии к неясным местам, дополненные при необходимости чтениями по другим спискам житий. Второй – электронное издание текста в формате XML в соответствии с рекомендациями Р4 консорциума Text Encoding Initiative (TEI). Отметим, что структура файлов проверена в редакторе XMLSPY, поэтому при необходимости они могут быть преобразованы в реляционные базы данных.

Опишем подробно XML-формат житийных текстов на сайте СКАТ.

Файл, описывающий структуру файлов (teiSCAT.dtd), приведен на странице сайта «XML-разметка текстов». В нем в виде условных обозначений (так называемых entities: &cyr-yat; &cyr-uk; и под.) указан перечень используемых древнерусских символов, отличных от стандартной кириллицы, поскольку не все древнерусские символы имеют надлежащее отображение в кодировке Юникод. При работе с текстами пользователям достаточно задать в файле teiSCAT.dtd привычные для них кодовые соответствия для символов «расширения», чтобы получить представление, полностью совместимое с локальной кодировкой в их исследовательских проектах.

В заголовке файлов приводится полная информация о выходных данных публикации серии «Памятники русской агиографической литературы», на базе которой создана электронная версия текста жития. Также указаны атрибуты того списка, который стал основой текста издания, и вспомогательных списков, используемых при исправлении ошибок и неясностей в основном тексте жития.

Формальное членение житийного текста осуществляется при помощи иерархической системы из пяти типов разделителей. Деление первого уровня (div0) – членение текста на собственно описание жития святого, чуда и похвальное слово. Как правило, житие и чудеса даются по одному списку, но в отдельных случаях (например, в Житии Павла Обнорского) тексты жития и чудес приводятся по двум разным рукописям. Разделители второго уровня (div1) показывают членение рукописи на листы, третьего (div2) – на лицевую и оборотную стороны рукописного листа, четвертого (div3) – на колонки, пятого (div4) – на рукописные строки.

Минимальным структурным элементом текста является слово (w), выделенное в процессе лингвистической интерпретации. Каждому слову приписывается идентификатор, состоящий из сокращенного обозначения конкретного жития в корпусе буквенно-цифровой последовательностью из двух-трех символов и порядкового номера слова в тексте. Сквозная нумерация элементов текста охватывает также и пунктуационные знаки.

Для описания слов предусмотрено несколько подполей-элементов.

Первый элемент (full) содержит подробный графематический состав слова: включены символы из расширения стандартной кириллицы, выносные буквы вставлены в линейную структуру слова, при этом их положение над строкой отмечено специальным атрибутом букв. Аналогично отмечено употребление покрытий над выносными буквами. Если слово в рукописном тексте помещается на двух (или более строках), то место «разрыва» помечено в линейной последовательности вставкой специального элемента (break), атрибут которого показывает тип «разрыва» – конец строки, колонки или листа. Если слово ошибочно (в нашей интерпретации), то в этом поле вставляется специальный атрибут (erroneous).

Второй элемент (reg) задает стандартное представление слова для поиска в словоуказателе. В настоящее время оно включает упрощение графематического состава до набора, включающего стандартную кириллицу без *й* и *э* с добавлением 'ять'. Имена собственные помечены специальным символом (*). Выносные буквы находятся на нужном по смыслу месте в слове и заключены в круглые скобки. Слова под титлом помечены в конце специальным знаком (#). В дальнейшем это поле может заполняться так, чтобы обеспечивать более удобный поиск, например, каноническим представлением слова (леммой в церковно-славянской или современной русской традиции). Для ошибочных написаний в этом поле приводится правильное чтение.

Третье поле (source) имеет служебный характер, оно показывает представление слова в кодировке DOS, которую мы используем с самого начала проекта.

На странице «Словоуказатель» сайта СКАТ есть возможность производить поиск по всем рукописям, подготовленным в виде электронных изданий. В качестве условия поиска задается последовательность не менее чем из трех букв, при этом можно указать положение буквосочетания в начале/середине/конце слова или задать словоформу полностью. Результаты сгруппированы по формам в алфавитном порядке, даны отсылки к тексту житий – указывается стандартное сокращение жития (полный перечень с расшифровками приведен при входе в поиск на странице «Словоуказатель»), номер листа, колонки, строки. При активизации ссылки вызывается контекстное окно, в котором отображается фрагмент pdf-файла – содержимое того рукописного листа, который указан в ссылке.

Список литературы

Аверина, С. А. Лексика и словообразование в русской агиографической литературе XVI в. : Опыт автоматического анализа / С. А. Аверина, И. В. Азарова, Е. Л. Алексеева, А. С. Герд ; под ред. А. С. Герда. – СПб., 1993.

- Аверина, С. А.* Лексика и морфология в русской агиографической литературе XVI в. / С. А. Аверина, И. В. Азарова, Е. Л. Алексеева, А. С. Герд, Л. А. Захарова, А. Д. Кривоносов ; под ред. А. С. Герда. – СПб., 1996.
- Аверина, С. А.* Язык русской агиографии XVI в. : Опыт автоматического анализа / С. А. Аверина, И. В. Азарова, Е. Л. Кузнецова и др. ; под ред. А.С. Герда. – Л., 1990.
- Азарова, И. В.* Разметка текстовых фрагментов в корпусе агиографических текстов СКАТ / И. В. Азарова, Е. Л. Алексеева, Л. А. Захарова // «Корпусная лингвистика 2006» : Тр. междунар. конф., 10-14 октября 2006 г. – СПб., 2006. – С. 16-24.
- Алексеева, Е. Л.* Разметка корпуса древнерусских текстов / Е. Л. Алексеева, А. М. Лаврентьев, И. В. Азарова, Л. А. Захарова // «Корпусная лингвистика 2004» : Тр. междунар. конф., 11-14 октября 2004 г. – СПб., 2004. – С. 16-24.
- Герд, А. С.* Корпус древнерусских агиографических текстов СКАТ: современное состояние и перспективы развития / А. С. Герд, И. В. Азарова, Е. Л. Алексеева, Е. С. Иванова // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам : материалы междунар. научн. конф. – Ижевск, 2006. – С. 38-42.
- Герд, А. С.* Электронный корпус текстов по памятникам древнерусской агиографической литературы / А. С. Герд, Е. Л. Алексеева, И. В. Азарова, Л. А. Захарова // НТИ. Сер. 2. Вып. 9. 2004. С. 16-20.
- Житие Кирилла Белозерского* : Текст и словоуказатель / Сост. И. В. Азарова, Е. Л. Алексеева, Д. Г. Демидов, Л. А. Захарова, М. Б. Попов ; под ред. А. С. Герда. – СПб., 2000.
- Житие Александра Свирского* : Текст и словоуказатель / Сост. И. В. Азарова, Е. Л. Алексеева, Л. А. Захарова, К. Н. Лемешев ; под ред. А. С. Герда. – СПб., 2002.
- Житие Антония Сийского* : Текст и словоуказатель / Сост. И. В. Азарова, Е. Л. Алексеева, Д. Г. Демидов, Л. А. Захарова, А. В. Сизиков ; под ред. А. С. Герда. – СПб., 2003.

Житие Кирилла Новоезерского : Текст и словоуказатель / Сост. И. В. Азарова, Е. Л. Алексеева, Л. А. Захарова, К. Н. Лемешев ; под ред. А. С. Герда. – СПб., 2003.

Жития Димитрия Прилуцкого, Дионисия Глушицкого и Григория Пельшемского : Тексты и словоуказатель / Сост. И. В. Азарова, Е. Л. Алексеева, Л. А. Захарова, К. Н. Лемешев ; под ред. А. С. Герда. – СПб., 2003.

Житие Корнилия Комельского : Текст и словоуказатель / Сост. И. В. Азарова, Е. Л. Алексеева, Л. А. Захарова, К. Н. Лемешев ; под ред. А. С. Герда. – СПб., 2004.

Жития Павла Обнорского и Сергия Нуромского : Тексты и словоуказатель / Сост. И. В. Азарова, Е. Л. Алексеева, Л. А. Захарова, К. Н. Лемешев ; под ред. А. С. Герда. – СПб., 2005.

Жития Иоасафа Каменского, Александра Куштского и Евфимия Сянжемского : Тексты и словоуказатель / Сост. И. В. Азарова, Е. Л. Алексеева, Л. А. Захарова, К. Н. Лемешев ; под ред. А. С. Герда. – СПб., 2007.

The St. Petersburg Corpus of Hagiographic Texts (SKAT): an XML format for encoding linguistic information and the site's search interface

Irina V. Azarova, Elena L. Alexeeva

St. Petersburg State University, St. Petersburg, Russia

The St. Petersburg Corpus of Hagiographic Texts contains 52 Russian texts (500,000 words) from the 16th and 17th centuries. So far 13 texts have been published online in PDF and XML formats. Texts are encoded according to TEI guidelines (P4). Search by stemming is available, and the query syntax allows entire word forms or letter combinations and their position (beginning, middle, or end) to be specified.