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

Корпус агиографических текстов СКАТ: XML-разметка элементов содержания

Е. А. Рогозина

XML-разметка, корпус, агиографические тексты, структура содержания, сюжет

The SCAT Digital Corpus of Hagiographic Texts: XML Encoding of Content Structure

Elena Rogozina

SCAT is a digital corpus of Old Church Slavonic hagiographic texts of the 15th–17th centuries created by the Department of Applied Linguistics of St. Petersburg State University. The texts are encoded in XML according to a formal division by pages, lines and words. At the same time it is possible to divided the texts to reflect their content structure since they were written in accordance with certain rules. Encoding of this structure facilitates search and further analysis of the texts.

СКАТ — электронный корпус агиографических церковнославянских текстов XVI–XVII вв., созданный на кафедре математической лингвистики СПбГУ. Работа над ним началась еще в 70-х годах XX века с создания картотеки житий святых русской церкви. Тогда же, в конце 70-х, началась работа по вводу текстов житий в компьютер. На данный момент корпус содержит более 50 житий общим объемом свыше 500 тысяч словоупотреблений. Одновременно с формированием базы данных было начато изучение грамматики и словообразования конкретных текстов. (Подробнее об истории проекта см. [Азарова, Алексеева, Захарова 2006: 16–24].)

С конца 90-х годов XX в. на кафедре математической лингвистики СПбГУ реализуется проект по изданию серии текстов “Памятники русской агиографической литературы”. В каждой книге помимо текстов одного или нескольких связанных между собой житий также представлен указатель словоформ, текстологический очерк и исторические сведения о святых. На данный момент опубликовано уже 11 выпусков серии, содержащих 23 жития вологодских святых.

Тексты опубликованных житий размещаются на сайте проекта[1] и доступны для бесплатного просмотра и скачивания. Помимо текстов на сайте представлена информация о самом проекте: история его создания, имена участников, принци-

[1] Санкт-Петербургский корпус агиографических текстов СКАТ // URL: http://project.phil.pu.ru/skat/

пы представления рукописного текста, особенности представления словоформ в словоуказателе.

Тексты житий представлены в формате PDF и формате XML. Разработка XML-формата для представления текстов корпуса началась на кафедре в 2004 году. Разметка осуществляется на основе международных норм оформления электронных изданий текста, в частности Text Encoding Initiative (TEI)[2].

На сайте проекта также представлен электронный словоуказатель, который позволяет осуществлять поиск словоформ по всему корпусу текстов, при этом можно искать как словоформу целиком, так и ее фрагмент. Для найденных словоформ указывается адрес: рукопись, номер листа и строки. Этот адрес соответствует формальному членению текстов на листы – колонки – строки – слова, которое отражено в структуре XML-файлов. Ведется также работа по разметке морфологических и синтаксических характеристик текстов.

Помимо формального деления на страницы и строки можно провести также деление на смысловые части. Прежде всего, следует отметить, что во многих рукописях автор сам разбивает текст на главы, выделяет их заголовками или буквицами. Разметив эти заголовки, можно создать для каждого жития оглавление по авторским разделам.

Однако для эффективного поиска по текстам этого деления недостаточно. В одной авторской главе может оказаться несколько различных сюжетов. Например, в авторской главе, отмеченной заголовком “О пострижении святого”, рассказывается не только о пострижении, но и о переходах святого из монастыря в монастырь, о желании удалиться в пустынь, о создании собственной обители и т. п. Поэтому возникла необходимость ввести более дробное смысловое деление.

Это возможно, поскольку композиция житий подчиняется определенным канонам. При написании жития автор ориентировался на уже существующие тексты. Готовый текст использовался как шаблон, который заполнялся новыми образами, персонажами, элементами сюжета [Панченко 2003: 491–534]. За счет этого приема каноническая схема жития заимствовалась и переходила из текста в текст. Поэтому можно вывести общую сюжетную схему, характерную для большинства из житий.

По результатам исследования текстов корпуса СКАТ была выделена общая схема развития сюжета, состоящая из характерных для всех исследуемых житий мотивов. Эта схема состоит из трех уровней. На первом уровне выделяются самые крупные элементы — блоки, которые описывают основную линию сюжета. На втором уровне каждый из блоков подразделяется на более мелкие компоненты. К третьему уровню относятся подвижные модули, которые могут встречаться на стыке или внутри элементов первого и второго уровней.

[2] Международный консорциум по выработке норм электронной разметки текстов [Электронный ресурс]. Режим доступа: http://www.tei-c.org/, свободный.

Не обязательно все выделенные блоки и компоненты присутствуют в каждом житии. Некоторые элементы схемы могут быть пропущены (например, может отсутствовать информация о жизни святого до пострига), другие же, наоборот, могут повторяться несколько раз (например, если святой переходил из одного монастыря в другой, повторяется блок с рассказом о жизни в монастыре).

Реализацию общей сюжетной схемы в конкретном тексте можно использовать в качестве оглавления, а XML-разметка позволяет указать для каждого блока или компонента точный адрес начала раздела и дать ссылку на соответствующую страницу текста.

Разметка содержательной структуры и создание оглавлений упрощает поиск по тексту, а также предоставляет дополнительные возможности для дальнейшего анализа, сопоставления блоков и компонентов в разных житиях. Такой анализ поможет выявить характерные для агиографических текстов обороты и цитаты, а также определить, есть ли закономерность в их употреблении.

Литература

Азарова, Алексеева, Захарова 2006 — *Азарова И. В., Алексеева Е. Л., Захарова Л. А.* Разметка текстовых фрагментов в корпусе агиографических текстов СКАТ // Труды междунар. конф. “Корпусная лингвистика–2006” (10–14 октября 2006 г.). СПб., 2006. С. 16–24.

Панченко 2003 — *Панченко О. В.* Поэтика уподоблений (к вопросу о “типологическом” методе в древнерусской агиографии, эпидейктике, гимнографии) // ТОДЛР. СПб., 2003. Т. 54. С. 491–534.