РАЗМЕТКА КОРПУСА ДРЕВНЕРУССКИХ АГИОГРАФИЧЕСКИХ ТЕКСТОВ

Алексеева Е.Л., Лаврентьев А.М., Азарова И.В.,Захарова Л.Д.

Корпус агиографических церковнославянских текстов XVI–XVII вв. на кафедре математической лингвистики Санкт-Петербургского государственного университета начал создаваться в конце 70-х годов. Житийный язык во многом обусловил судьбу и характер русского литературного языка той эпохи, поскольку в это время – XVI–XVII вв. – на Руси создавалось множество оригинальных житий русских святых со своими особенностями и в содержании, и в языке.

Работа началась с создания картотеки житий святых русской церкви, похвальных слов, сказаний, в которой учитывались исследования и издания этих текстов; были изысканы средства для образования фонда фото- и ксерокопий рукописей житий, находящихся в разных рукописных хранилищах Петербурга, который постоянно пополняется. Тогда же, в начале 70-х, началась работа по вводу текстов житий в компьютер.

Параллельно формированию базы данных было начато изучение грамматики, словообразования конкретных текстов. В результате к концу 1996 г. вышло в свет три обобщающих книги, которые содержат систематическое описание именного склонения, глагольного спряжения и именного словообразования памятников русской агиографической литературы XVI вв., опубликован ряд словоуказателей, полученных на ЭВМ [1].

Наконец, с конца 90-х годов XX в. на кафедре математической лингвистики СПбГУ реализуется широкомасштабный проект по изданию уникальной серии текстов «Памятники русской агиографической литературы». Каждое такое издание содержит текст жития и полный словоуказатель словоформ, а также вводные статьи по истории текста, краткую биографию святого, сведения об обителях[2].

---

[1] *Аверина С. А., Азарова И. В., Кузнецова Е. Л. и др.* Язык русской агиографии XVI в.: Опыт автоматического анализа / Под ред. А.С. Герда. Л., 1990 (здесь же см. и литературу по языку русской агиографии); *Аверина С. А., Азарова И. В., Алексеева Е. Л., Герд А. С.* Лексика и словообразование в русской агиографической литературе XVI в.: Опыт автоматического анализа / Под ред. А.С. Герда. СПб., 1993; *Аверина С. А., Азарова И. В., Алексеева Е. Л., Герд А. С., Захарова Л. А., Кривоносов А. Д.* Лексика и морфология в русской агиографической литературе XVI в. / Под ред. А. С. Герда. СПб., 1996.

[2] *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*: Текст и словоуказатель / Сост. И. В. Азарова, Е. Л. Алексеева, Л. А. Захарова, К. Н. Лемешев; Под ред. А. С. Герда. СПб., 2003.

К настоящему времени корпус охватывает 52 жития, их общий объем – более 500 тыс словоупотреблений.

По сравнению с текстами на современном языке тексты XVI в. обладают рядом особенностей, которые существенно затрудняют и замедляют процесс создания электронного корпуса: во-первых, текст не разбит на слова; во-вторых, текст, вообще говоря, нелинеен, поскольку в нем широко используются выносные буквы, иногда окончание предыдущей строки занимает правую позицию в следующей строке, нередко встречаются вставки с полей; в-третьих, это рукописный текст, т.е. он может содержать ошибочные написания, как неисправленные, так и исправленные; в-четвертых, одна и та же буква может быть представлена несколькими графическими вариантами, которые могут иметь, а могут и не иметь особого значения. Поэтому представление текста жития в электронном виде предполагает решение нескольких содержательных задач – определение границ словоразделов, представление в линейном виде нелинейного текста, выявление ошибочных написаний слов (в отличие от графических вариантов написания одного и того же слова), интерпретация используемых в рукописях различных написаний букв и диакритических знаков и т. п.

Для представления рукописей в корпусе была разработана система отображения древнерусской графики, которая позволяет воспроизводить текст с высокой степенью приближения к оригиналу. Отображены графические начертания всех древнерусских букв и их семантически значимых вариантов (узкое и широкое «о»; узкое, широкое, якорное «е» и т. п.). Воспроизводятся титла, титловые покрытия, паерки, выносные буквы и буквосочетания, а также знаки придыхания и акцентные знаки. Разработка базового шрифта для ввода житийных текстов представляла собой ряд последовательных приближений к выявлению набора необходимых и достаточных знаков, при этом не преследовалась цель фототипической точности воспроизведения рукописей, например, варианты букв, не имеющие фонетического или палеографического значения, лигатуры, «лежачие» начертания выносных букв в базе житийных текстов не отображаются.

Разработана специальная программа, позволяющая получать к введенным текстам (к каждому в отдельности или к нескольким вместе) указатели словоформ, т.е. списки словоформ с их адресами (номерами листов и строк) в рукописях.

Здесь нам также пришлось столкнуться с некоторыми проблемами. В алфавите, который Древняя Русь восприняла от южных славян, уже с самого начала были буквы, не имевшие особого фонетического значения, например, в нем было 3 буквы для звука И, 2 буквы для О, 2 буквы для Ф и т.д. К XVI в. некоторые буквы поменяли свое звуковое значение уже на русской почве, в языке развились такие фонетические явления, как аканье, позиционное оглу-

шение и озвончение шумных согласных, все это привело к тому, что одна и та же словоформа могла быть записана несколькими способами. К тому же писцы в своей работе очень часто использовали приемы сокращенного написания слов (под титлом или с выносными буквами), и в текстах житий некоторые словоформы имеют свыше 10 вариантов написания.

Таким образом при написании программы составления словоуказателя нужно было решать проблему сведения графических вариантов словоформ к одному виду.

Во-первых, словоформы в словоуказателе представлены в упрощенной графике: устранено дублирование букв, опущены надстрочные знаки, выносные буквы в круглых скобках спущены в строку на свое место по смыслу.

Во-вторых, особая программа[3] осуществляет частичное сведение орфографических вариантов одной и той же словоформы к единой форме:

— объединяются словоформы с одинаковым буквенным составом, различающиеся наличием/отсутствием выносных букв или тем, какие именно буквы помещены над строкой, например: **безмо(л)вии** – **бе(з)молвии** – **безмолвии** ⇒ **безмолвии**; **чадоро(ди)ю** – **ча-(до)родию** – **чадородию** ⇒ **чадородию**;

— объединяются словоформы, различающиеся тем, как представлен конечный согласный: выносная буква/ строчная буква без ера или еря/ строчная буква с ером или ерем, например: **игумe(н)** – **игумен** – **игуменъ** ⇒ **игуменъ**;

— объединяются длинные (шесть и более букв) словоформы под титлом, если одна из них длиннее/короче на одну букву другой, например: **блженнаго**# – **блженаго**# – **блжнаго**# ⇒ **блженаго**#;

— объединяются словоформы, различающиеся тем, как представлена частица **же** в их составе: полностью или в виде выносного **ж**, например: **поне(ж)** – **понеже** ⇒ **понеже**;

— объединяются словоформы, различающиеся тем, как представлено возвратное **ся** в их составе: полностью или в виде выносного **с**, например: **повину(с)** – **повинуся** ⇒ **повинуся**;

— объединяются (в диалоговом режиме) словоформы, различающиеся видом редуцированного на конце слова, например: **старецъ** – **старець** ⇒ **старець**;

[3] Программа составлена выпускниками Санкт-Петербургского государственного университета Тарасовой Еленой Евгеньевной и Тарасовым Евгением Анатольевичем.

— объединяются (в диалоговом режиме) словоформы, если букве **а** в одной соответствует **о** в другой: **манастыря** – **монастыря** ⇒ **монастыря**; **студенаго** – **студеного** ⇒ **студенаго**; **алтаря** – **олтаря** ⇒ **алтаря**.

Эта программа позволяет уменьшить объем словника на 6–10%.

В корпусе тексты житий представлены дважды – в текстовом формате и формате редактора Word. В текстовом файле отсутствуют диакритические знаки и выносные буквы в круглых скобках вставлены на свое место в слове по смыслу, на основании текстового файла создается словоуказатель. Текст в редакторе Word также создается на основе текстового файла, в него вносятся диакритические знаки и все выносные буквы занимают свое место над строкой, по внешнему виду этот текст приближается к тексту рукописи, но имеет существенное отличие: он разделен на слова.

В настоящее время полнотекстовый формат корпуса дополняется XML разметкой, структурированно отображающей компоненты смысловой и формальной интерпретации текста, которые были выявлены в ходе обработки достаточно большого количества текстов и были сформулированы в публикациях коллектива составителей корпуса. А.М.Лаврентьев написал программу для автоматической конвертации текстового файла в формат XML. Основной целью этой части проекта является представление опубликованных текстов в виде, доступном для внешних пользователей. Эти данные будут представлены на сайте филологического факультета СПбГУ в режиме электронного доступа.

В формальном плане разметка корпуса основывается на международных нормах оформления электронных изданий текста, в частности Text Encoding Initiative[4] (TEI), однако содержит также и дополнительные элементы, которые необходимы для адекватного отображения особенностей русского рукописного текста.

По схеме TEI заполняется «паспорт» электронного варианта рукописи, который включает следующие части. (1) Выходные данные оригинального текста рукописи, который был использован в корпусе как основной. Помимо этой обязательной информации, мы предусматриваем также указание на принадлежность текста жития к определенной редакции; выходные данные других рукописей этой редакции, которые используются для прояснения «темных» мест в тексте. (2) Следующий блок информации — кодировка текста. Типичное решение в этом разделе – указание кодов символов из древнерусской части системы Юникод. К

сожалению, этот раздел не включает все символы, которые необходимы для представления житийных текстов. В частности нет нескольких основных символов (йотированного «а», широкого «е», «ы» в виде сочетания «ера» с «и десятиричным»), а также тех вариантов начертания букв, которые имеют палеографическое значение. (3) В следующем разделе указывается автор текста жития, при этом приводится отсылка к тексту рукописи, либо к историческим источникам, на основании которых установлено авторство. Кроме того, факультативно приводится информация о писцах (если эти данные есть в рукописи). (4) Заключительный раздел описания документа включает перечень авторов-составителей, которые отвечают за принятие смысловых решений при интерпретации рукописного текста: проведение словоразделов, выявлении ошибочных написаний и проч.

В основу структуры жития как электронного документа положены формальные характеристики рукописи: разбивка текста на листы, колонки, строки. Эта информация представлена и в текстовом файле, она автоматически перекодируется в тэги (метки) начала/конца листа, колонки, строки с соответствующей нумерацией.

Уже отмечались такие особенности средневековых рукописных текстов, как отсутствие словоразделов, не всегда линейное представление текста. Представляя текст в электронном формате, можно выбрать один из двух путей: максимально точно воспроизводить вид рукописного текста, а его смысловую интерпретацию приводить в качестве меток, или наоборот: воспроизводить текст, а особенности его представления отмечать метками. Мы предпочли второй путь. Точно так же, когда формальное членение текста (на строки и листы) не совпадает со смысловым (на слова), мы всегда сохраняем целостность текста, то есть слово, перенесенное с одной строки на другую, представляется не в виде двух раздельных элементов, а целиком, но при этом отмечается место, где проходит граница строки.

В нашей разметке мы используем определенный набор атрибутов.

Один из атрибутов характеризует размер букв, который определяется не столько реальными физическими параметрами написания буквы, сколько соотношением со смысловым делением рукописи на части: инициал, заглавная буква и строчная буква. Еще один атрибут отмечает использование киновари. Особый атрибут предусмотрен для имеющихся в рукописи исправлений.

Что касается слов, отмечаются имена собственные и ошибочные написания, при этом в особых тэгах приводится «правильное» написание, которое либо дано редакторами транс-

[4] Международный консорциум по выработке норм электронной разметки текстов // URL: http://www.tei-c.org/P4X/

крипции рукописи, либо зафиксировано в рукописях, которые были рассмотрены как параллельные поясняющие тексты.

Из фрагментов особым образом отмечаются, в первую очередь, заголовки (они регулярно написаны киноварью, вынесены в отдельную строку и проч.), а также предусматривается выделение библейских цитат со ссылкой на соответствующее место в Библии.

Верифицированные xml-представления житийных текстов будут в дальнейшем дополнены морфологической разметкой: фрагменты житий уже размечены вручную, однако метки еще не переведены в электронный вид.