1.1.2.1 Исправление, инициированное партнером
1.1.2.2 Проверка контакта
1.1.2.3 Уточнение условий ответа (VTE VTJ)
1.2 Информационные акты
1.2.1 Директивы (DIE DIJ)
1.2.2 Вопросы (KYE KYJ)
KYE: общий вопрос
KYE: общий вопрос, ожидающий развернутого ответа
KYE: альтернативный вопрос
KYE: специальный вопрос
KYE: иное
KYJ: да
KYJ: нет
KYJ: согласное нет
KYJ: иной ответ на общий вопрос
KYJ: альтернатива: одна
KYJ: альтернатива: обе
KYJ: альтернатива: третий выбор
KYJ: альтернатива: отрицание
KYJ: альтернатива: иное
KYJ: развернутый ответ
KYJ: отсутствие информации
KYJ: отказ
KYJ: иное
1.2.3 Мнение
II Одиночные акты
2.1 Акты управления диалогом
2.1.1 Коммуникация
2.1.1.1 Ритуалы (RY)
2.1.1.2 Обратная связь
2.1.2 Разрешение проблем
2.1.2.1 Исправление
2.2 Информационные акты
2.2.1 Основные акты (YA)
2.2.2 Дополнения основных актов (пояснение, уточнение)

Примечание: Подробно расписана группа вопросов и ответов — наиболее частотных речевых актов; эстонская аббревиатура, предшествующая названию акта, содержит информацию о группе и о позиции акта в смежной паре.

Корпус древнерусских агиографических текстов СКАТ: современное состояние и перспективы развития

А. С. Герд, И. В. Азарова, Е. Л. Алексеева, Е. С. Иванова
Санкт-Петербургский государственный университет, Россия

The Corpus of Russian hagiographic texts of the 16–17th centuries at present comprises 52 texts or 500 000 word-tokens, represented in 2 formats: as text files and Microsoft Word files; the texts are provided with a word form index. 10 texts have been published; they are available to Internet users in the PDF and XML formats. The work is under way to provide all texts with the morphological information. Automatic normalization of varying Church-Slavonic spelling is another important task.

Корпус агиографических церковнославянских текстов XVI–XVII вв. на кафедре математической лингвистики Санкт-Петербургского государственного университета начал создаваться в конце 70-х годов. 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 конце 70-х, началась работа по вводу текстов житий в компьютер. К настоящему времени корпус охватывает 52 жития, их общий объем — более 500 тыс. словоупотреблений.

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

словоуказателей, полученных на ЭВМ [Опыт 1990; Лексика 1993; Лексика 1996].

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[1].

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, то есть списки словоформ с их адресами (номерами листов и строк) в рукописях.

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 раз-

[1] В 2000–2006 гг. опубликованы: Житие Кирилла Белозерского, Житие Александра Свирского, Житие Антония Сийского, Житие Кирилла Новоезерского, Жития Димитрия Прилуцкого, Дионисия Глушицкого и Григория Пельшемского, Житие Корнилия Комельского, Жития Павла Обнорского и Сергия Нуромского.

вились такие фонетические явления, как аканье, позиционное оглушение и озвончение шумных согласных, отвердение некоторых исконно мягких 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, становится очевидной актуальность проблемы сведения графических вариантов словоформ к одному виду.

Для решения этой проблемы мы используем несколько приемов. Во-первых, упрощение графики: устранено дублирование букв, опущены акцентные знаки, выносные буквы в круглых скобках спущены в строку на свое место по смыслу; объединяются словоформы с одинаковым буквенным составом, различающиеся наличием/отсутствием выносных букв или тем, какие именно буквы помещены над строкой. Во-вторых, восстановление полного буквенного состава словоформ, пишущихся в сокращенной форме: восстанавливаются до полного вида корни, регулярно сокращаемые под титлом[2]; объединяются словоформы, различающиеся тем, как представлен конечный согласный: выносная буква — строчная буква с редуцированным — строчная буква без редуцированного); объединяются словоформы, различающиеся тем, как представлена частица ЖЕ или возвратное СЯ в их составе: полностью или в виде, соответственно, выносного Ж или С. В-третьих, устранение графического варьирования, являющегося следствием изменений фонетической системы языка: унифицируется написание гласных с шипящими согласными и Ц, с заднеязычными согласными; устраняется варьирование А/О в начале слова и в окончании -ОГО/-АГО; объединяются словоформы, различающиеся наличием или отсутствием интервокального йота; приводится к одному виду написание корней с плавными сонантами в сочетании с редуцированными; унифицируется написание некоторых морфем, в составе которых исходно имелся редуцированный гласный.

В настоящее время мы осуществляем грамматическую разметку представленных в корпусе житий. Разработан формат пред-

[2] Мы различаем титло и покрытие, ставящееся над выносной буквой.

ставления грамматической информации для всех частей речи, который в виде цифрового кода вносится в текстовый файл. Первая цифра кода означает часть речи, интерпретация остальных цифр зависит от того, к какой части речи относится слово.

В корпусе тексты житий представлены дважды — в текстовом формате и формате редактора Word. В текстовом файле предусмотрена кодировка для всех букв славянского алфавита, текст рукописи поделен на слова и представлен в линейном виде: выносные буквы в круглых скобках вставлены на свое место в слове по смыслу, опущены акцентные знаки. Особыми символами отмечаются концы строк, столбцов и листов рукописи; границы текста, вносимого с полей; ошибочные написания. 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, отличаясь от него тем, что он разделен на слова.

Нами было принято решение сделать опубликованные тексты житий доступными для внешних пользователей и представить их на сайте филологического факультета СПбГУ, для этого текст в формате Word преобразуется в формат PDF, а текстовый файл — в формат XML, конвертация осуществляется автоматически.

В формальном плане XML-разметка корпуса основывается на международных нормах оформления электронных изданий текста, в частности Text Encoding Initiative (TEI) [3].

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.

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 особенно-

[3] Международный консорциум по выработке норм электронной разметки текстов: The Text Encoding Initiative [Электронный ресурс]. — Режим доступа: http://www.tei-c.org/P4X/, свободный. — Загл. с экрана.

сти 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.

Верифицированные XML-представления житийных текстов будут в дальнейшем дополнены морфологической разметкой: разработаны формат представления грамматической информации и программа автоматической конвертации грамматических кодов в тэги формата XML.

Список литературы

Аверина и др. 1990 — *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.

Аверина и др. 1993 — *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 и др. 1996 — *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 [и др.] ; под ред. А. С. Герда. — СПб., 1996.