Е. Л. Алексеева, И. В. Азарова
E. L. Alexeeva, I. V. Azarova

ОСОБЕННОСТИ МОРФО-СИНТАКСИЧЕСКОЙ РАЗМЕТКИ ДРЕВНЕРУССКИХ АГИОГРАФИЧЕСКИХ ТЕКСТОВ

PECULIARITIES OF THE MORPHO-SYNTACTIC ANNOTATION FOR OLD RUSSIAN HAGIOGRAPHIC TEXTS

Аннотация. В докладе рассматриваются особенности морфо-синтаксической разметки в Санкт-Петербургском корпусе агиографических текстов (СКАТ). Помимо стандартной спецификации частеречных тегов и значений грамматических категорий, аннотация отмечает варьирование типов основ, приведших к смешению типов склонения существительных, формирование категории одушевленности, утраты двойственного числа, перестройку системы глагольных форм прошедшего времени. На синтаксическом уровне отмечаются связи синтаксической зависимости с полным и неполным согласованием.

Abstract. Saint-Petersburg Corpus of Hagiographic Russian texts (SCAT) uses a pilot model for the morpho-syntactic text annotation. It includes a standard set of POS tags and their grammatical values, moreover, special markers highlight the peculiar shifts in the system of Old Russian: alteration of stem types leading to the confusion of declension types, development of noun animateness, loss of the dual, reorganization of verbal past forms. The syntactic tagging in stand-off files allows to represent syntactic dependencies with full and limited coordination matching.

1. Санкт-Петербургский корпус агиографических текстов

На кафедре математической лингвистики Санкт-Петербургского государственного университета идет работа по проекту СКАТ, формированию корпуса русских агиографических текстов XV–XVII вв. В настоящее время в корпус входит более 65 рукописей, общим объемом более полумиллиона словоупотреб-

лений. Тексты в корпусе прошли через трудоемкую процедуру подготовки[1]: предварительного анализа графемного состава рукописи, деления текста на слова, что предполагает морфосинтаксический и семантический анализ текста, представления текста рукописи в электронной форме. На каждом из этапов возникают вопросы, которые решаются исследовательским коллективом СКАТ. Результат работы над житийными текстами представлены в публикациях[2], которые содержат тексты рукописей с подстрочными примечаниями, комментирующими неясные места (они приводятся в тексте в оригинальном виде), что позволяет увидеть нашу трактовку смысла текста. На сайте СКАТ опубликованные рукописи доступны для общего пользования в виде pdf-файлов. Кроме того, тексты рукописей представлены в xml-формате, который позволяет преобразовать их при необходимости в другую кодировку и другую систему тегов.

2. Представление рукописных текстов в xml-файлах

Базой для представления рукописных текстов в формате xml служит их материальный носитель, который определяет разбивку рукописи на части (собственно житие и похвальное слово), пронумерованные листы с пометой лицевой и оборотной сторо-

---

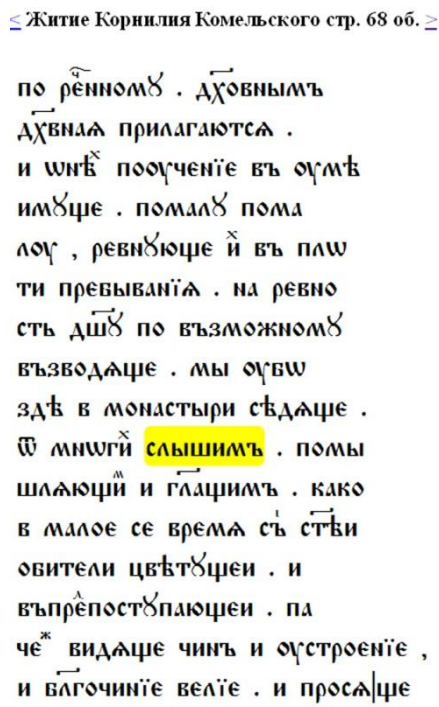
[1] *Герд А.С., Алексеева Е.Л., Азарова И.В., Захарова Л.А.* Электронный корпус текстов по памятникам древнерусской агиографической литературы // НТИ. Сер. 2. Вып. 9. 2004. С. 16–20; *Герд А.С., Азарова И.В., Алексеева Е.Л., Иванова Е.С.* Корпус древнерусских агиографических текстов СКАТ: современное состояние и перспективы развития // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам. Ижевск, 2006. С. 38–42.

[2] В 2012 г. вышел 11 выпуск в серии «Памятники русской агиографической литературы»: Жития Феодосия Тотемского, Вассиана Тиксненского и Андрея Тотемского (под ред. А.С. Герда; СПб.: Изд-во С.-Петерб. ун-та, 2012). Там же приведен перечень опубликованных житий.

ны, пронумерованные колонки текста (если таковые есть в рукописи) и пронумерованные строки.

Рис. 1. Фрагмент Жития Корнилия Комельского (л. 68 об.), преобразованный из xml-файла на сайте **http://scat.v-alexeev.ru**

Эта система разделителей позволяет при обратном преобразовании воспроизвести внешний вид рукописи, как



показано на рис. 1.

В базовых xml-файлах воспроизводится графемный состав рукописи на том уровне, который коллектив СКАТ счел информативным. Каждое выделенное слово рукописи снабжено числовым идентификатором, что позволяет однозначно определить

вхождение слова в текст. Знаки препинания выделяются в качестве отдельных элементов (c). Помимо полного графемного представления xml-файл содержит представление слова в упрощенной графике, которое используется при поиске в словоуказателе по корпусу. На рис. 2 приведен фрагмент xml-файла, описывающий начало текста, показанного на рис.1.

```
<div2 type='page' n='68'>
        <div3 type='back'>
                <div4 type='col' n='1'>
                        <l n='1'>
<w xml:id='KrnKml.70'>
        <orig>по</orig>
        <reg>ПО</reg>
        <src>ПО</src>
</w>
<w xml:id='KrnKml.71'>
        <orig>рѣ□нном□</orig>
        <reg>РЕ(ч)НЮМУ</reg>
        <src>РЕ(ч)НЮMD</src>
</w>
<c type='punctuation' xml:id='KrnKml.72'>.</c>
<w xml:id='KrnKml.73'>
        <orig>дх҃овнымъ</orig>
        <reg>ДХОВНЫМЪ#</reg>
        <src>ДХОВНЫМЪ#</src>
</w>
                        </l>
                        <l n='2'>
<w xml:id='KrnKml.74'>
        <orig>дх҃внаѧ</orig>
        <reg>ДХВНАЯ#</reg>
        <src>ДХВНАR#</src>
</w>
```

Рис. 2. Фрагмент xml-файла для Жития Корнилия Комельского (л. 68 об.) на сайте **http://project.phil.spbu.ru/scat/**

В случае ошибок в тексте рукописи приводится исходный вариант написания слова и исправленный вид, как показано в (1).

(1) <w xml:id='KrnKml.3810'>

```
        <orig><choice><sic>забвевена</sic><corr>забвена</corr></choice></orig>
        <reg>~ЗАБВЕВЕНА &lt;ЗАБВЕНА&gt;</reg>
        <src>~ЗАБВЕВЕНА &lt;ЗАБВЕНА&gt;</src>
</w>
```

3. Морфологическая разметка рукописных текстов

В настоящее время морфологическая разметка житийных текстов проводится вручную, в рамках лингвистической практики студентов, и затем выверяется квалифицированным специалистом коллектива СКАТ. Основным препятствием для автоматической разметки является неустойчивая орфография, которая приводит к высокой вариативности написания слов. Таким образом, морфологическая разметка текстов является однозначной. В дальнейшем размеченные тексты будут использоваться в качестве прецедентной совокупности для автоматических процедур.

Морфологическая разметка приводится в отдельном файле, поскольку идентификаторы слов (см. выше: *xml:id='KrnKml.71'*) позволяют однозначно привязывать аннотацию к элементу текста. При необходимости файл разметки и базовый текстовый файл могут быть объединены в один.

Набор тегов морфологической аннотации[3] включает стандартный набор тегов: спецификация частей речи (2) и соответствующие наборы значений морфологических категорий (3).

(2)
```
<fsdDecl>
  <fsDecl type="scatGramData">
    <fsDescr>Grammatical properties of words in the SCAT corpus</fsDescr>
    <fDecl name="POS">
      <fDescr>Part of speech</fDescr>
      <vRange>
        <vAlt>
          <symbol value="noun"/>
          <symbol value="pronoun"/>
```

[3] Спецификация морфо-синтаксической разметки следует рекомендациям TEI P5: Guidelines for Electronic Text Encoding and Interchange. Eds. L. Burnard, S. Bauman 2007. URL: **http://www.tei-c.org/ release/ doc/tei-p5-doc/en/Guidelines.pdf**

```
      <symbol value="adjective"/>
      <symbol value="cardinal numeral"/>
      <symbol value="participle"/>
      <symbol value="verb"/>
      <symbol value="infinitive"/>
      <symbol value="supine"/>
      <symbol value="adverb"/>
      <symbol value="preposition"/>
      <symbol value="postposition"/>
      <symbol value="conjunction"/>
      <symbol value="particle"/>
      <symbol value="interjection"/>
    </vAlt>
  </vRange>
 </fDecl>
</fsDecl>
</fsdDecl>
```

Помимо частеречной спецификации используется субкатегоризация классов, например, прилагательное / краткое, прилагательное / сравнительная степень, инфинитив / возвратная форма и т. д.

Подклассы частей речи и значения грамматических категорий определяются в виде структуры значений признаков (3). В примере показано, что семь значений категории падежа образуют систему взаимоисключающих значений.

(3)
```
<fDecl name="case">
        <fDescr>Case</fDescr>
        <vAlt>
                <symbol value="nominative"/>
                <symbol value="genitive"/>
                <symbol value="dative"/>
                <symbol value="accusative"/>
                <symbol value="ablative"/>
                <symbol value="locative"/>
                <symbol value="vocative"/>
        </vAlt>
</fDecl>
```

4. Синтаксическая разметка рукописных текстов

Синтаксическая разметка в TEI P5 предполагает выделение синтаксических групп различных уровней: предложений, слово-

сочетаний и проч. Однако для житийных текстов пунктуационное выделение не является достаточным основанием для проведения четких границ предложений. В большей степени пунктуация похожа на синтагматическое членение текста, когда выделяются синтагмы и фразы. В любом случае этот вопрос носит дискуссионный характер. Поэтому мы предполагаем начать с достаточно ясных отношений синтаксической зависимости, когда можно выделить главный и зависимый элементы, между которыми будем фиксировать полное или неполное согласование.

В нашем модельном исследовании будем использовать структуру зависимостей наиболее простого типа (4).

(4) <**graph**
type="directed"
xml:id="DepStructure_KrnKml"
...>
<**arc** from="#KrnKml.75" to="#KrnKml.76">
<**label**>CoordFull</**label**>
</**arc**>
<**arc** from="#KrnKml.75" to="#KrnKml.79">
<**label**>NoCoord</**label**>
</**arc**>
</**graph**>

5. Особенности морфо-синтаксической разметки древнерусских текстов

Наиболее сложной задачей, на наш взгляд, является отражение переходных явлений, которые присутствуют в житийных текстах. Архаические формы аориста встречаются наряду с новыми, происходит смешение типов склонения существительных, утрачивается система склонения кратких прилагательных, развивается категория одушевленности в системе склонения существительных, реорганизуется система прошедших времен глаголов, утрачивается парадигма двойственного числа. Игнорировать эти процессы и просто фиксировать «новую» форму, например, винительного, совпадающую с родительным, было бы недальновидно. Более того, такого рода пометы необходимы для

исследовательских задач. К сожалению, информация такого типа не предусмотрена в руководстве TEI P5, но мы предполагаем использовать схему, сходную с корректировкой ошибок в тексте, указывая значение грамматической категории в тексте и ее интерпретацию.