И.В. Азарова, Е.Л. Алексеева, В.А. Алексеев

АВТОМАТИЗИРОВАННАЯ ГРАММАТИЧЕСКАЯ РАЗМЕТКА КОРПУСА АГИОГРАФИЧЕСКИХ ТЕКСТОВ СКАТ

Формирование агиографического корпуса СКАТ началось почти 30 лет тому назад, и к настоящему времени объем корпуса достиг 500 тыс. словоупотреблений. 13 из 50 введенных житий XVI–XVII вв. опубликованы и доступны пользователям интернета на сайте **http://www.project.phil.pu.ru/scat** в двух форматах: PDF и XML.

На странице «Словоуказатель» сайта СКАТ можно производить поиск словоформ по представленным на нем рукописям, задавая в качестве условия поиска словоформу целиком или последовательность из трех или более букв с указанием положения буквосочетания в слове: в начале, середине или в конце.

Для обеспечения более релевантных результатов автоматического поиска для запрашиваемых слов вводится морфологическая разметка словоформ в корпусе.

В 2006 году Е.С. Ивановой был разработан формат грамматической разметки[1], в дальнейшем уточненный Е.Л. Алексеевой, рассчитанный на ручную разметку текстов. В нем число заполняемых позиций для изменяемых частей речи варьируется от 4 до 6, указываются: часть речи, для именных частей речи – тип склонения, род, число, падеж, для глагольных – наклонение, время, лицо, число, класс спряжения (для форм настоящего времени и повелительного наклонения) и т.п. Особым образом размечаются словоформы в составе сложных форм глаголов.

Не все значения категорий для данной словоформы задаются в формате эксплицитно: в некоторых случаях определенное зна-

[1] *Иванова Е.С.* Схема разметки текста для электронной публикации древнерусских рукописей. Дипломное сочинение, рукопись. СПб., 2006.

чение одной категории однозначно задает значение другой, и мы это учитываем. Например, если местоимение является возвратным, не нужно указывать число (всегда единственное), у причастий тип склонения коррелирует с залогом: действительные причастия склоняются по мягкой парадигме, страдательные – по твердой. Программа обработки заполненных таблиц разметки в таких случаях самостоятельно определяет не заданное вручную значение соответствующей категории.

В соответствии с этим форматом были полностью размечены вручную два текста корпуса общим объемом 23 тыс. словоупотреблений. Полученный таким образом материал используется далее для разработки автоматизированной системы грамматической разметки агиографических текстов.

В.А. Алексеевым разработан алгоритм и написана программа для определения формы и леммы простых форм глаголов изъявительного наклонения, при условии предварительного определения части речи и наклонения.

В процессе работы над автоматизацией разметки глаголов мы столкнулись со следующими проблемами:

1. Необходимость учета палатализации согласных (чередования заднеязычных с шипящими и т.п.), возникающей в положении перед *j*.
2. Наличие двух параллельно встречающихся форм образования имперфекта (со стяжением и без).
3. Невозможность однозначного определения тематической гласной для форм прошедшего времени, а также в ряде случаев и для настоящего времени.
4. Наличие глаголов с нерегулярной парадигмой.
5. Существование омонимичных форм.
6. Орфографическая вариативность написания словоформ.
7. Сокращенное написание слов под титлом.

Первые две проблемы решаются программным путем.

Определение тематической гласной в сложных случаях и обработка неправильных глаголов не представляются возможным без участия человека. Когда программа находит словоформу,

которая может иметь несколько вариантов леммы или морфологической характеристики, она просит пользователя выбрать правильный вариант и записывает его в файл обучения; когда программа встречается с подобным выбором в дальнейшем, она берет ответ из этого файла. Формы, образующиеся нерегулярным образом, вручную записываются в специальный файл, где указывается лемма и грамматические показатели такой формы.

Снятие омонимичности форм возможно только вручную на этапе, предшествующем автоматической обработке словоформ. Имеются следующие случаи омонимии:

1. Омонимия второго и третьего лица аориста. По умолчанию все слова автоматически размечаются как имеющие форму третьего лица, как более часто встречающуюся. Случаи второго лица отмечаются вручную.
2. Омонимия второго лица множественного числа настоящего-будущего времени с третьим лицом двойственного числа. По умолчанию слова размечаются как принадлежащие второму лицу. Случаи третьего лица отмечаются вручную.
3. Омонимия первого лица аориста и имперфекта. В связи с тем, что их различение в большой мере субъективно, было принято решение не различать эту омонимию, поэтому все слова размечаются, как имеющие форму аориста.

Для сведения к одному виду графических вариантов словоформ и раскрытия форм под титлом используется процедура, написанная Е.Г. Уфлянд.

Создание процедуры автоматизированного определения форм и лемм глаголов позволило упростить формат ручной разметки простых форм глаголов: вместо 5–6 позиций обязательно заполняются две: частеречная принадлежность и наклонение, и в относительно небольшом числе случаев дополнительно указывается лицо глагола для снятия омонимии.

На данном экспериментальном материале программа надежно определяет форму глагола, но процедура определения леммы требует доработки.