Е.Г. Уфлянд, Е.Л. Алексеева

СОКРАЩЕНИЕ ВАРИАТИВНОСТИ НАПИСАНИЯ СЛОВОФОРМ В СЛУЖЕБНЫХ КОМПОНЕНТАХ АГИОГРАФИЧЕСКОГО КОРПУСА СКАТ

Агиографический корпус СКАТ в настоящее время включает около 50 древнерусских житий XVI–XVII вв., причем 13 опубликованных текстов доступны пользователям интернета на сайте **http://www.project.phil.pu.ru/scat** в двух форматах: PDF и XML.

На странице «Словоуказатель» сайта СКАТ предусмотрена возможность производить поиск словоформ по представленным на нем рукописям. Базой для поиска является указатель словоформ, содержащий для каждой словоформы адреса всех ее вхождений (сигла рукописи, номер листа, номер колонки (если листы рукописи поделены на колонки), номер строки). В качестве условия поиска можно задать словоформу целиком или последовательность из трех или более букв, дополнительно указав положение буквосочетания в слове: в начале, середине или в конце.

Как известно, в России вплоть до 1917 г. не существовало четко установленных правил правописания, на написание любой конкретной словоформы могли влиять несколько факторов: орфография древних рукописей, живая произносительная норма, второе южнославянское влияние, использование сокращения под титлом или с выносными буквами. Это приводило к тому, что одна и та же словоформа могла иногда иметь более 10 различных вариантов написания, например: **блаженаго** – **блаженна** – **блаженнаго** – **блаенннаго** – **блаанна** – **блааннаго** – **блжена** – **блженаго** – **блженаго** – **блаженна** – **блаженнаго** – **блжна** – **блжнаго** – **блжнна** – **блжна** – **блжнаго** – **блжннаго**. Очевидно, что подобный графический разнобой существенно снижает эффективность поиска.

Нами была поставлена задача проанализировать факторы, обусловливающие вариативность орфографии агиографических памятников XVI–XVII вв., выделить регулярно встречающиеся модели вариативности и реализовать автоматическое сведение вариантов словоформы к одному основному варианту. В качестве основного варианта, как правило, принимается наиболее близкий к современному написанию.

Е.Г. Уфлянд была написана процедура на языке Python, позволяющая устранить следующие случаи графической вариативности:

- сокращенные написания сакральных слов под титлом: **мчнк**, **мнк** > **мученик**;
- вынос отдельных букв или буквосочетаний над строкой: **грешна** , **грешна, грешнаго** > **грешнаго**;
- написание «ъ/ь» после заднеязычных согласных: **отвергьше**, **отвергъше**, **отвергше** > **отвергше**;
- написание «ъ/ь» после «м» и «в»: **мьзду**, **мзду** > **мзду**;
- написание «ъ/ь» после шипящих и «ц»: **мужъ**, **мужь** > **мужꙗ** ;
- написание финального «ъ»: **нас**, **насъ** > **насъ**;
- написание сочетаний плавного и редуцированного между согласными в корне: **длъжни**, **дльжни**, **должни** > **должни**;
- написание гласных после шипящих: **одержымъ**, **одержимъ** > **одержимъ**;
- написание гласных после заднеязычных согласных: **книгы**, **книги** > **книги**;
- написание сочетаний гласных: **въсвоаси**, **въсвоꙗси** > **въсвоꙗси**;
- написание суффиксов «-ьск-», «-ьств-»: **богатество**, **богатьство**, **богатство** > **богатство**;

- написание «и/ь» перед последней гласной окончания:
  келии, **кельи** > **келии**.

Применение этой процедуры к словоуказателю позволило сократить его объем почти на 20%. На данном материале процедура работает без ошибок, но можно предположить, что расширение материала приведет к появлению ошибочных сведений вариантов словоформ, поэтому все произведенные программой замены записываются во вспомогательный файл, что позволяет контролировать правильность работы программы.

Однако нельзя сказать, что вариативность в словоуказателе устранена полностью. Нам предстоит решить еще ряд проблем, важнейшими среди которых являются

а) достаточно многочисленные случаи чередования редуцированных и гласных «о/е» на месте современных беглых гласных в приставке и корне: **въсхитити** – **восхитити**,

б) удвоенные согласные: **воистинѵ** – **воистиннѵ**,

в) непоследовательное написание ятя: **грѣхъ** – **грєхъ**.

Устранение орфографического разнобоя имеет большое значение не только для поиска словоформ в представленных на сайте текстах, решение этой проблемы должно существенно облегчить еще одну задачу, стоящую перед создателями корпуса – задачу автоматизированной грамматической разметки и лемматизации.