ATLANTIS PRESS
*Advances in Social Science, Education and Humanities Research (ASSEHR), volume 122*
45th International Philological Conference (IPC 2016)

# Принципы нормализации древнерусских житийных текстов для корпуса СКАТ

**Азарова Ирина Владимировна[1] Алексеева Елена Леонидовна[2]**

**Сипунин Константин Владимирович[3]**

[1]Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург, Россия
[2]Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург, Россия
[3]Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург, Россия

# Normalization of Old Russian Hagiographic Texts in SCAT

**Azarova Irina[1] Alekseeva Elena[2] Sipunin Konstantin[3]**

[1]Saint-Petersburg State University (SPSU), St. Petersburg, Russia
[2]Saint-Petersburg State University (SPSU), St. Petersburg, Russia
[3]Saint-Petersburg State University (SPSU), St. Petersburg, Russia

**Аннотация**

На кафедре математической лингвистики СПбГУ создан и постоянно пополняется корпус агиографических текстов (СКАТ), в котором представлены тексты древнерусских житий по рукописям XVI–XVIII вв. Доклад посвящен проблемам нормализации орфографии и лемматизации в словоуказателе к текстам житий.

**Abstract**

Department of Mathematical Linguistisc of SPSU maintains the corpus of hagiographic texts (SCAT), comprising digital versions of Church Slavonic handwritten lives of saints from 16-18th centuries. The paper deals with lemmatization and spelling normalization in the index to the texts.

**Ключевые слова**: СКАТ, исторический корпус, нормализация орфографии, лемматизация

**Keywords**: SCAT, historical corpus, spelling normalization, lemmatization

Санкт-Петербургский корпус агиографических текстов (СКАТ) создается для представления электронных интерпретаций древнерусских житийных текстов XV-XVII вв. исследователями кафедры математической лингвистики СПбГУ совместно со специалистами по

 

текстологии ИРЛИ РАН (Пушкинский дом), библиотеки Российской академии наук (БАН) и др. учреждений. В корпус помещаются подготовленные к изданию тексты Житий, при этом сохраняется орфография конкретной рукописи, выбранной в качестве представителя данного текста, а неизбежные в каждой рукописи описки и смысловые неясности проясняются по другим спискам жития [Герд 2006]. Адрес сайта http://project.phil.spbu.ru/scat/page.php?page=project.

**Формат представления текстов и организация процедуры поиска**

В корпусе тексты представлены в двух форматах: pdf и xml.

В формате xml единицей описания является словоформа, знак пунктуации или числовое обозначение, которым присваиваются однозначные адреса (id), состоящие из сокращенного названия файла и порядкового номера элемента в тексте. Напр., в нижеследующем примере записи слова *игумена* в xml-файле идентификатор *DPrlc.12*означает 12-й элемент в «Житии Димитрия Прилуцкого»:

<wxml:id="DPrlc.12"><orig>игоу<lb/>мена</orig><reg>ИГУМЕНА</reg><src>ИГU&МЕНА</src></w>.

Слово в записи представлено в трех вариантах:

(1) элемент orig показывает, как слово выглядит в рукописи: сохранен исходныйсостав графем и обозначен разрыв строки (тег lb/), пришедшийся на середину слова;
(2) элемент reg дает нормализованное написание слова для его представления в словоуказателе (графема оу заменена на у, убран разрыв строки);
(3) элемент src— вид слова в текстовом файле, который является исходным для ряда программ обработки текста (графема *оу* обозначена *u*, разрыв строки — &).

Сводный указатель словоформ к житиям обеспечивает поиск слов по всем текстам как на основе строгого, так и нестрогого соответствия: можно искать как словоформу целиком, так и ее фрагмент. На запрос пользователя выдается список всех словоформ, удовлетворяющих условиям поиска, и по указанным адресам можно автоматически перейти к соответствующим фрагментам текста. Сейчас мы работаем над тем, чтобы уменьшить объем словника за счет нормализации написания словоформ и лемматизации.

**Нормализация орфографии словоформ**

Во всех исторических корпусах устранение вариативности написания слов является одной их первоочередных задач, поскольку от ее решения зависит эффективность организации поиска по корпусу, точность морфологической разметки и лемматизации.

М. Больманном (Bollmann)была предложена трехступенчатая процедура нормализации орфографии для немецкого языка, целью которой было установление соответствий между древними и современными формами [Bollmann 2013]:

(1) просматривается список соответствий древних и современных словоформ, полученный вручную на

обучающем корпусе объемом до 1000 единиц, и устанавливаются однозначные варианты;

(2) для словоформ, не получивших интерпретации на первом этапе, используется «нормализация по правилам»: позиционная замена одних символов другими;

(3) если замены на втором этапе не дали результата, используется взвешенное расстояние Левенштейна:каждой операции замены приписан вес, и всовременном лексиконе выбирается та словоформа, к которой приводит цепочка замен с наименьшей суммой весов.

Точность нормализации прямо пропорциональна объему обучающей выборки и при выборке в 1000 текстовых единиц для текстов XV в. составляет примерно 80%, а для XVII в. — 87–95%.

В рамках проекта СКАТ мы не ставим задачу поиска соответствий древним формам в современном языке, поскольку, например, для многих форм глаголов в результате упрощения системы времен за счет развития категории вида таких соответствий нет. Нашей задачей является унификация написания словоформ с предпочтением варианта, наиболее близкого к современному написанию.

В основе вариативности орфографии в древнерусских и церковнославянских текстах лежат как объективные причины: наличие дублетных графем в алфавите, фонетические процессы, использование сокращенных написаний слов, так и субъективные: индивидуальные особенности графики писца, особое написание слов на границе двух строк, ошибки и описки.

Учет ряда объективных закономерностей позволил создать программный модуль, уменьшающий список словоформ почти на 20% [Уфлянд 2008]. Программа отрабатывалась на материале 10 текстов житий общим объемом 120 579 словоупотреблений. Были разработаны контекстнозависимые и контекстносвободные правила замены буквосочетаний и выносных букв, и составлен словарь для раскрытия сокращенных написаний под титлом. Хотя показатель полноты сведения вариантов словоформ к единому виду составил всего 50%, это отчасти компенсируется высокой точностью: 99,7%. Дальнейшая работа планируется в двух направлениях: расширение списка правил для закономерных написаний словоформ и разработка процедуры оценки близости окказиональных написаний словоформ к имеющимся регулярным формам.

**Лемматизация**

На протяжении нескольких лет силами студентов кафедры математической лингвистики проводится ручная морфологическая разметка текстов житий корпуса [Азарова 2013]. На основании разметки существительных К.В. Сипуниным был разработан и программно реализован алгоритм определения леммы для каждой словоформы.

Для существительного в разметке приводятся следующие сведения: тип склонения с различением твердой или мягкой парадигмы, падеж, число и род.Также предусмотрена возможность отражения переходных явлений: через косую черту приводится ожидаемое значение соответствующей категории и реально встретившееся в тексте[Алексеев 2011]. Напр.:

- тип склонения *о/и* для существительного *доуховъ* обозначает, что оно относится к склонению на *-ŏ, но имеет окончание склонения на *-ŭ;
- падеж *вин/род* для существительного *игоумена* показывает, что в значении винительного падежа употреблен родительный — так проявляется категория одушевленности.

На первом этапе работы морфологическая характеристика каждого существительного была представлена в виде четырехзначного числового кода, содержащего номера типа склонения, падежа, числа и рода, и был составлен вспомогательный словарь окончаний, в котором каждому окончанию была поставлена в соответствие совокупность возможных числовых кодов. На основании этого словаря в словоформе отыскивается и отсекается флексия, а затем присоединяется окончание исходной формы соответствующего типа склонения. В программе учитываются релевантные фонетические явления, такие как пропуск и прояснение редуцированных гласных, смешение *е* и *ѣ*, употребление заднеязычных согласных перед гласными переднего ряда.

Программа была проверена на материале «Жития Димитрия Прилуцкого» и для 1329 существительных правильно определила лемму в 1296 случаях (97,52%). Анализ ошибок показал, что есть возможность добиться еще более высокого результата.

## Литература

*Азарова И. В., Алексеева Е. Л.*Особенности морфо-синтаксической разметки древнерусских агиографических текстов // Труды международной конференции «Корпусная лингвистика — 2013». СПб: СПбГУ, Филологический факультет, 2013. С. 157-164.

*Алексеев В.А., Алексеева Е.Л., Касьяненко С.Е.* Грамматическая разметка в корпусе СКАТ // Труды международной конференции «Корпусная лингвистика — 2011». СПб: СПбГУ, Филологический факультет, 2011. С. 69-73.

*ГердА.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 рукописей к электронным текстам. Материалы международной научной конференции. Ижевск: Изд-во ИжГТУ, 2006. С. 38-42.

*Уфлянд Е.Г., Алексеева Е.Л.* Сокращение вариативности написания словоформ в служебных компонентах агиографического корпуса СКАТ // Труды международной конференции «Корпусная лингвистика — 2008». СПб: СПбГУ, Факультет филологии и искусств, 2008. С. 376-378.

*Bollmann M.* POS Tagging for Historical Texts with Sparse Training Data // Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. Sofia, 2013. P. 11-18.

## References