Правительство Российской Федерации
Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Санкт-Петербургский государственный университет»

Выпускная квалификационная работа

на тему:

Автоматическая лемматизация текстов в корпусе СКАТ на основе морфологической разметки

Основная образовательная программа бакалавриата

по направлению подготовки 45.03.02 «Лингвистика»

Исполнитель:
обучающийся 4 курса
Образовательной программы «Прикладная,
экспериментальная и математическая
лингвистика (английский язык)»
Профиль «Прикладная, экспериментальная
и математическая лингвистика»
очной формы обучения
Сипунин Константин Владимирович

Научный руководитель:
к. ф. н., доц. Алексеева Е. Л.

Рецензент:
к. ф. н., доц. Захаров В. П.

Санкт-Петербург
2018

Аннотация

В данной выпускной квалификационной работе исследуется проблема разработки автоматизированных инструментов для лемматизации морфологически размеченных житий в составе Санкт-Петербургского корпуса агиографических текстов (СКАТ). В рамках теоретической части исследования производится обзор существующих восточнославянских исторических корпусов в аспекте реализованных в них технологий грамматической разметки. Практическая составляющая работы посвящена проблемам именного словоизменения в церковнославянском языке позднесредневекового извода и методам их формального учёта применительно к задаче лемматизации, а также организации полноценного доступа к корпусу СКАТ (включая его размеченный и лемматизированный сегмент) при помощи платформы TXM.

Ключевые слова: древнерусская агиография, грамматическая разметка, исторический корпус, церковнославянское словоизменение, электронное представление рукописей

Abstract

This graduation paper deals with the problem of developing automatic tools for lemmatizing morphologically annotated vitae comprising the Saint Petersburg Corpus of Hagiographic Texts (SCAT). As a theoretical background of the present work, a survey of existing East Slavic historical corpora is carried out, with special attention paid to the technological aspects of their grammatical annotation. The experimental part addresses the issues of nominal inflectional morphology in late medieval Church Slavonic and the procedures involved in their formalized solution as applied to the problem of lemmatization, as well as the provision of full access to SCAT (including its annotated and lemmatized subcorpus) by means of the TXM platform.

Keywords: Old Russian hagiography, grammatical annotation, historical corpus, inflection in Church Slavonic, digital representation of manuscripts

Оглавление

Введение

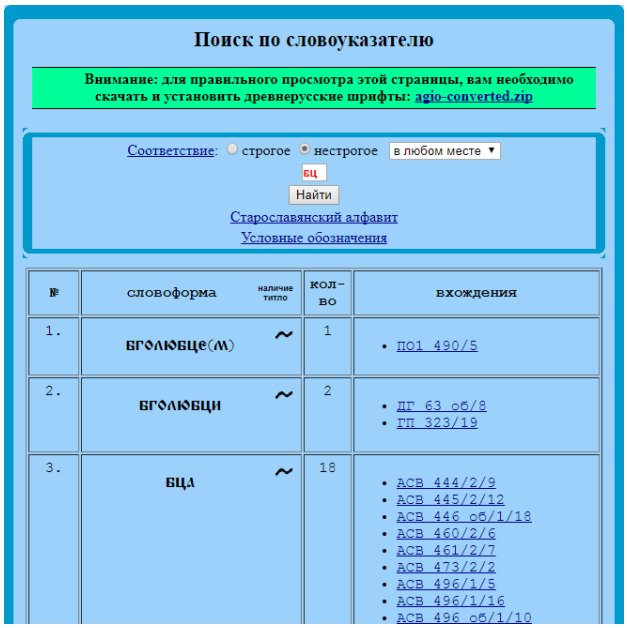
Славянская рукописная традиция зародилась уже более тысячи лет назад. От времён, последовавших за просветительской деятельностью преподобных Константина (Кирилла) и Мефодия в середине IX в., до сегодняшних дней дошли десятки тычяч рукописей, созданных писцами и переписчиками в монастырях Восточной Европы, — как списков Священного Писания, служебников, часословов и прочих богослужебных книг, непосредственно обслуживавших запросы церкви, так и оригинальных произведений, предназначенных для индивидуального чтения: поучений, сказаний, житий святых. Тем не менее, значительная доля данных текстовых массивов по сей день изучена недостаточно и по-прежнему нуждается во всесторонней исследовательской обработке — исторической, этнографической, лингвистической.

Несколько десятилетий назад ситуация начала качественно преображаться в связи с появлением, а впоследствии и массовым распространением компьютеров и цифровых технологий: средства представления рукописей в электронном виде ознаменовали собой принципиально новые возможности их сохранения и изучения вне стен отдельных библиотек и архивных фондов, регулярным доступом к которым обладает далеко не каждый исследователь.

Соответствующие разработки начали появляться уже в конце третьей четверти XX в. — в том числе на кафедре математической лингвистики Ленинградского государственного университета. С конца 1970-х гг. при участии сотрудников кафедры русского языка ЛГУ, а также ИРЛИ АН СССР и ГПБ им. М. Е. Салтыкова-Щедрина на кафедре начал создаваться фонд фото- и ксерокопий списков древнерусских житий и похвальных слов XV–XVII вв. [1, с. 512], впоследствии получивший название «Санкт-Петербургский корпус агиографических текстов» (СКАТ). Для представления содержимого фонда в памяти ЭВМ каждую копию рукописного текста было необходимо трансли-

терировать — перевести в машиночитаемый формат при помощи специальной системы кодирования. Однако в те годы фактически единственным средством ввода символьных цепочек в память компьютера являлись 8-битные кодировки на базе ASCII (American Standard Code for Information Interchange), очевидно не предназначенные для размещения в диапазоне кодируемых символов знаков устаревших и экзотических систем письменности (в т. ч. кириллической). Вследствие этого для набора текстов, составляющих фонд, на кафедре была выработана собственная кодировка, в которой для вышедших из употребления символов кириллицы были введены замены (преимущественно буквы латинского алфавита): так, юс большой и юс малый обозначаются соответственно «G» и «R», кси — «L» и т. д. Тексты вводимых в память ЭВМ рукописей набираются квалифицированными специалистами-филологами вручную при помощи специально разработанного шрифта AGIO и затем автоматически переводятся в данную кодировку; при этом в текст вставляются словоразделы (в соответствии с принципами, разработанными проф. А. А. Алексеевым для издания серии «Библиотека литературы Древней Руси»), а также маркируются границы составных частей рукописи — строк, колонок и страниц. Всего к настоящему времени в базу данных введено более полусотни рукописей общим объёмом около полумиллиона словоупотреблений [11].

Сегодня доступ к результатам работы коллектива проекта обеспечивается двояко. С одной стороны, с конца 1990-х гг. издательством Санкт-Петербургского государственного университета ведётся публикация изданий серии «Памятники русской агиографической литературы», в каждом из которых содержится один или несколько подготовленных к печати житийных текстов, набранных упомянутым выше шрифтом AGIO, полный словоуказатель словоформ, а также текстологические статьи об истории публикуемых житий, биографии святых, сведения об обителях. Последний, одиннадцатый выпуск



(*а*) Выдача по запросу бц (режим нестрогого соответствия)

(*б*) Контекстное окно вхождения словоформы бголюбци# (ГП 323/19)

Рис. 1. Поиск по словоуказателю СКАТ

увидел свет в 2012 г.; там же приведён перечень всех предыдущих публикаций серии [3, с. 4].

С другой стороны, всё более повсеместное распространение онлайн-технологий в 2000-х гг. дало импульс к тому, чтобы обеспечить доступ к опубликованным материалам через интернет: был создан сайт проекта[1], а корпус получил своё нынешнее наименование. На сегодняшний день около полутора десятков житий доступны для загрузки с сайта в двух форматах: PDF, воспроизводящем их представление в печатных сборниках, и XML, где с помощью системы тегов производится формальное членение рукописей на структурные элементы. XML-разметка текстов СКАТ соответствует международному стандарту оформления электронных изданий — Text Encoding Initiative (TEI).

Также на сайте имеется возможность поиска по корпусу — вернее, по той его части, для которой построен сводный словоуказатель. Это центральный компонент лингвистического обеспечения СКАТ, представляющий со-

[1] http://project.phil.spbu.ru/scat/ (дата обр. 01.06.2018)

бой список словарных статей, в каждой из которых указана словоформа в нормализованном виде, абсолютная частота её встречаемости по всем проиндексированным рукописям и адреса вхождений. Адрес состоит из сокращённого наименования рукописи и сочетания порядковых номеров листа (с уточнением стороны — лицевой либо оборотной), колонки и строки, разделённых косой чертой. При нажатии на адрес в поисковой выдаче (рис. 1, *а*) пользователю предлагается «нарезка» из соответствующего PDF-документа (рис. 1, *б*), в которую попадает искомое вхождение; отыскивать его приходится самостоятельно — путём отсчитывания от межстраничной либо межколонной границы с номером, указанным в адресе, необходимого числа строк.

Однако современный электронный корпус — в отличие от простой коллекции текстов — должен располагать определённым набором автоматизированных инструментов, применимых в ходе решения конкретных лингвистических задач. В ряде зарубежных работ по языкам с ограниченными ресурсами в последние годы вошло в обиход понятие BLARK — Basic Language Resource Toolkit (базовый набор лингвистических ресурсов), которое определяется как «the minimal set of language resources that is necessary to do any precompetitive research and education at all» [27, p. 11] (минимальный набор лингвистических ресурсов, необходимый для любых базовых исследовательских и образовательных нужд). BLARK может включать в себя как традиционные одно- и двуязычные словари и грамматики, так и специфические ресурсы, вошедшие в лингвистический обиход лишь в последние десятилетия: модули распознавания и синтеза речи, морфосинтаксические анализаторы и т. д. Притом отмечается, что этот список не закрытый и может варьироваться от языка к языку: очевидно, для древнеписьменных языков, в число которых входит и церковнославянский, неактуальна задача обработки устной речи, однако вследствие некодифицированного характера орфографии зачастую требуются модули её нормализации.

Passarotti [32, p. 28] предлагает вариант BLARK («a BLARK-like set») для латинского языка, который, как кажется, в равной степени приложим к другим древнеписьменным языкам. В нём предусмотрены инструменты, направленные на решение следующих основных задач: (1) предобработка текстовых данных: токенизация и распознавание именованных сущностей; (2) морфологический анализ: лемматизация и разрешение морфосинтаксической неоднозначности; (3) синтаксический анализ (поверхностный и глубинный); (4) разрешение анафоры; (5) семантический и прагматический анализ.

Целью настоящей работы является разработка комплекса инструментов для осуществления процедуры лемматизации словоформ именных частей речи в составе церковнославянских текстов корпуса СКАТ. **Задачи**, которые необходимо решить для достижения поставленной цели, таковы:

1. ознакомление с системами представления грамматической информации (в т. ч. данных по леммам) в существующих восточнославянских исторических корпусах;
2. изучение теоретических вопросов именного словоизменения в церковнославянском языке XV–XVII вв. с целью их формального учёта в ходе программной разработки алгоритма лемматизации;
3. организация доступа к лемматизированному подкорпусу СКАТ (и шире — ко всей оцифрованной части корпуса) с использованием общедоступных технологических средств.

Объект основной части исследования — именное словоизменение в церковнославянском языке XV–XVII вв., т. е. словоизменение имён существительных, прилагательных, числительных, а также местоимений. **Предмет** изучения — проблемы формализации феноменов церковнославянского именного словоизменения в ходе алгоритмизации перехода от словоформ в несловарных парадигматических позициях к словарным формам (леммам). **Материалом** послужили морфологически размеченные тексты трёх агиогра-

фических текстов в составе корпуса СКАТ: жития Димитрия Прилуцкого, Дионисия Глушицкого и Кирилла Новоезерского — суммарным объёмом около 30 тыс. словоупотреблений.

Актуальность работы обоснована тем, что в рамках СКАТ — единственного в своём роде источника сведений по языку древнерусской агиографии эпохи позднего Средневековья и Нового времени — серьёзные попытки разработки составных частей BLARK в целом и подсистем морфологического анализа в частности фактически не предпринимались.

Структура работы включает в себя введение, 3 главы, заключение, список литературы из 35 наименований и 2 приложения.

Глава 1

Представление грамматической информации в восточнославянских исторических корпусах

Список восточнославянских исторических корпусов, приведённый на сайте Национального корпуса русского языка[2] (НКРЯ), включает в себя следующие наименования (помимо СКАТ):

1. Регенсбургский диахронический корпус русского языка[3];
2. Рукописные памятники Древней Руси[4];
3. система «Манускрипт»[5];
4. корпус русских публицистических текстов второй половины XIX в.[6]

Среди перечисленных корпусов первый и третий, а также СКАТ, исторические подкорпуса самого НКРЯ и корпус «Великие Минеи Четьи»[7] обзорно освещены в статье [31]: каждый охарактеризован с точки зрения характера текстов, лежащих в их основе, суммарного объёма (актуального на момент написания статьи), возможностей поиска, наличия морфосинтаксической аннотации и прочих релевантных признаков. В данной главе мы рассмотрим описанные О. В. Митрениной корпуса в аспекте реализованных в них принципов и инструментов грамматической разметки и лемматизации несколько более детально (исключая лишь корпус «Великие Минеи Четьи», в котором последние отсутствуют [35, p. 30]).

---

[2] http://ruscorpora.ru/corpora-other.html (дата обр. 01.06.2018)

[3] http://rhssl1.uni-regensburg.de/SlavKo/korpus/rrudi-new (дата обр. 01.06.2018)

[4] http://www.lrc-lib.ru (дата обр. 01.06.2018) Отметим, что базы древнерусских берестяных грамот и летописей, входящие в архив данного ресурса, к настоящему времени полностью интегрированы в древнерусский сегмент НКРЯ и потому в особом рассмотрении не нуждаются.

[5] http://manuscripts.ru (дата обр. 01.06.2018)

[6] http://smalt.karelia.ru/corpus/index.phtml (дата обр. 01.06.2018) Ввиду своей временной специфики он выбивается из общего ряда восточнославянских исторических корпусов и поэтому останется за рамками настоящего обзора.

[7] http://www.vmc.uni-freiburg.de/Mens/ (дата обр. 01.06.2018)

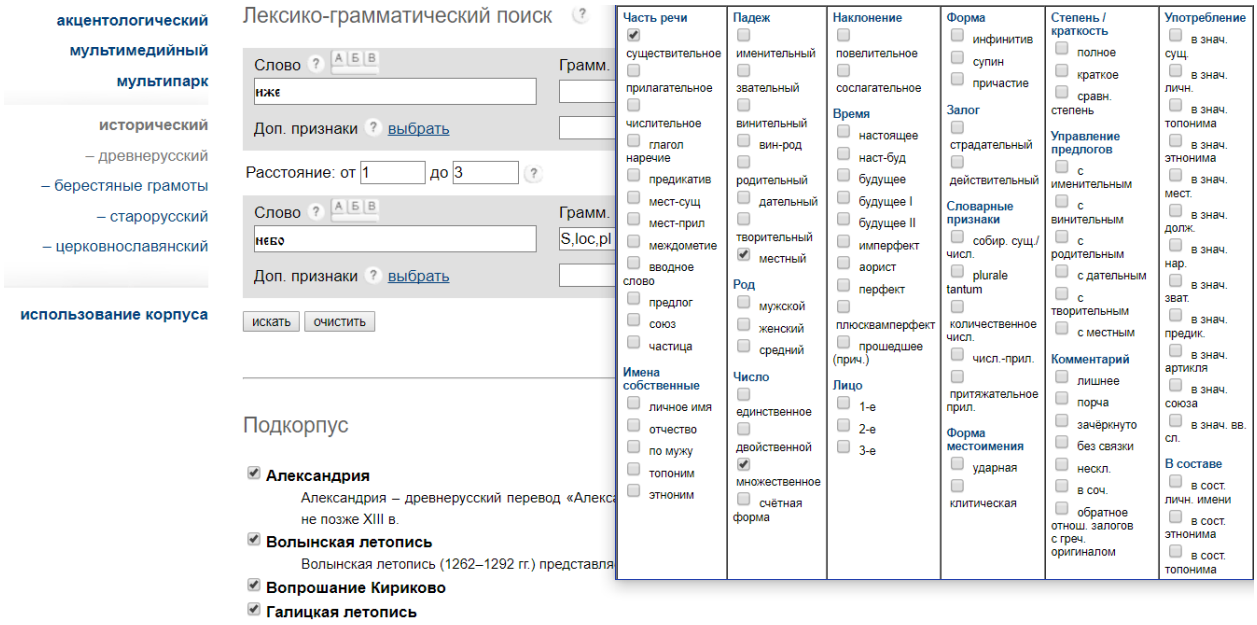
1.1. Национальный корпус русского языка

1.1.1. Древнерусский корпус и корпус берестяных грамот

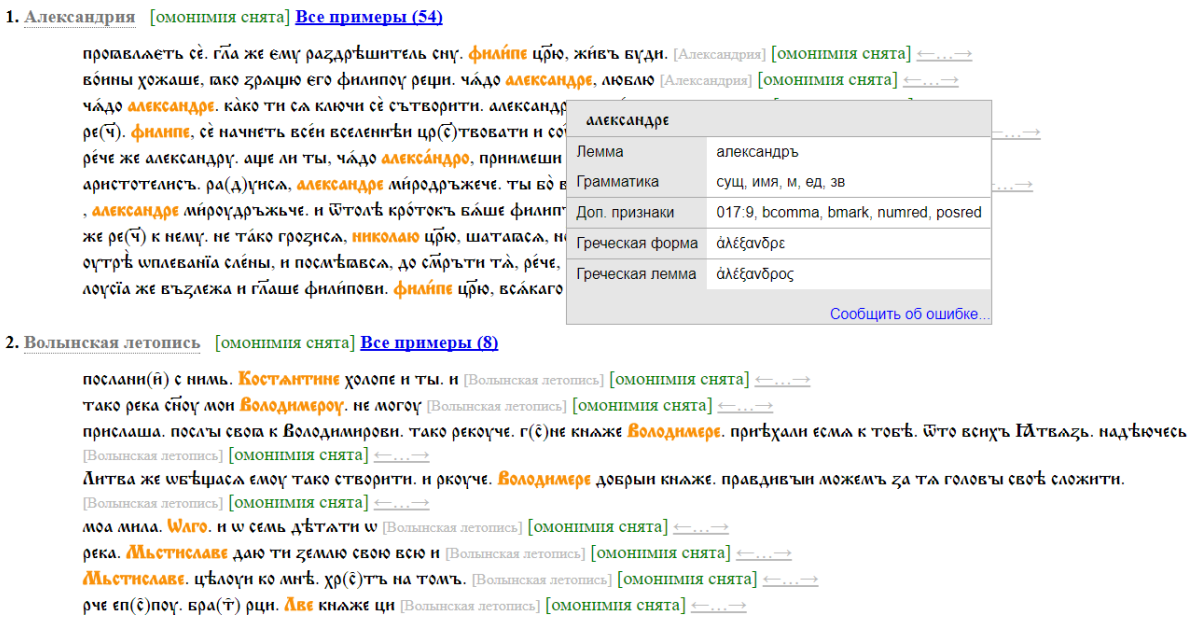
Эти подкорпуса НКРЯ будут рассмотрены совместно: их разработка ведётся по единой программе президиума РАН «Корпусная лингвистика», а составляющие их тексты написаны на одном (пусть и значительно неоднородном) языке и датируются сходными хронологическими периодами — XI–XIV и XI–XV вв. соответственно. Грамматическая разметка текстов, объединённых в древнерусский корпус, ведётся ещё с середины 2000-х гг. в рамках проекта «Рукописные памятники Древней Руси»; содержимое же корпуса берестяных грамот основывается на материалах сборников «Новгородские грамоты на бересте» и размечается для сайта «Древнерусские берестяные грамоты»[8] (в настоящее время доступ к разметке с него не предоставляется) [20, с. 226].

Разметка обоих корпусов производится вручную со снятием грамматической омонимии. При аннотации древнерусского корпуса ввиду его существенно большего объёма (на сегодняшний день он составляет около 500 тыс. словоупотреблений — в противовес 20 тыс. в случае корпуса берестяных грамот) производится дополнительное обращение к базе прецедентных разборов [17, с. 102]; аппарат разметки корпуса берестяных грамот по сравнению с древнерусским «усовершенствован в связи с большим количеством фрагментарно сохранившихся слов, а также мест, трактуемых лишь предположительно» [20, с. 227], однако в остальном оба корпуса размечены в соответствии с едиными принципами. Решение собственно технических задач ввода разметки, построения словоуказателей, работы со словарями и прочими компонентами лингвистического обеспечения осуществляется разработчиками посредством специальной среды Morphy [7].

[8] http://gramoty.ru (дата обр. 01.06.2018)



(*а*) Поисковый интерфейс



(*б*) Выдача по запросу на личные имена в звательном падеже

Рис. 1.1. Древнерусский подкорпус НКРЯ

Помимо повсеместных граммем, актуальных отнюдь не только для древнеписьменных языков (часть речи; падеж, число, род; наклонение, время, залог и т. д.), рассматриваемые корпуса располагают средствами разметки специфически древнерусских морфологических характеристик: например, для личных местоимений в дат. и вин. п. предусмотрены специальные пометы для ударных и клитических форм. Кроме того, в формат аннотации словоформ включены элементы традиционной филологической адресации по номерам листа и строки, а также ряд помет принципиально иного рода, например ономастических (имена собственные могут размечаться как личные имена и отчества, обозначения жены или вдовы по мужу, топонимы и этнонимы) и текстологических (для мест с неясной интерпретацией введены пометы «зачёркнуто», «лишнее», «порча»). Запросная форма и диалоговый интерфейс выбора грамматических признаков проиллюстрированы на рис. 1.1, *а*.

Разработчики отмечают, что «разбор древнерусского текста — работа, не вполне поддающаяся стандартизации. … Поскольку потребность добавлять релевантные пометы возникает при изучении древних текстов постоянно, система разметки разрабатывалась таким образом, чтобы исследователь мог легко вводить новые признаки по собственному усмотрению» [17, с. 103–104]. Как следствие, разные памятники в своей аннотации порой обнаруживают известную степень неоднородности, обусловленной личными предпочтениями и взглядами исследователей на размечаемые лингвистические феномены: в частности, подобные расхождения проявляются при разборе имён собственных (в одних памятниках соответствующие пометы проставляются, в других же разметчики ограничиваются базовой частеречной типизацией) и форм аналитических будущих времён, по сей день недостаточно изученных и неоднозначно трактуемых в исторической русистике [17, с. 104–106].

В отдельных переводных памятниках (в частности, это касается «Александрии» и «Пчелы», воспроизводимых в корпусе по спискам XV в.) сло-

воформам по возможности приписаны греческие аналоги — причём как сами эквивалентные формы, так и их леммы (см. контекстное окно при существительном *АЛЕКСАНДРЪ[9] на рис. 1.1, *б*). Собственно же древнерусские леммы составляют общий для всех памятников словарь, где они представлены в унифицированной (не содержащей дублетных графем) древнерусской орфографии, отражающей состояние до падения и прояснения редуцированных [17, с. 102–103]. К сожалению, суммарный объём словаря разработчиками не уточняется.

1.1.2. Церковнославянский корпус

Церковнославянский подкорпус НКРЯ охватывает лишь те тексты на церковнославянском языке, которые были созданы или отредактированы уже в период книгопечатания — в XVII–XX вв., причём основная доля (60 %) приходится на современные тексты, т. е. используемые в современной богослужебной практике. Этим обстоятельством объясняется объём корпуса, по меркам исторических корпусов весьма внушительный: его составляют более 1250 документов, охватывающих все основные типы и жанры церковнославянской литературы и включающих более 4,7 млн словоупотреблений, которые группируются в 150 тыс. различных словоформ [18, с. 246–247]. Очевидно, ручной морфологический анализ столь обширных текстовых массивов принципиально нереализуем.

Грамматическая аннотация церковнославянского корпуса одновременно нацелена как на процедуру лемматизации — приведение словоформы к лемме и определение её постоянных признаков (части речи, рода, вида, переходности), так и на собственно грамматический анализ — выявление грамма-

[9] Здесь и далее языковые примеры даются капителью с соблюдением транслитерационных соглашений, принятых в коллективе СКАТ: вышедшие из употребления графемы обозначаются при помощи символов латиницы; октоторп маркирует наличие в слове титла; выносные буквы заключаются в скобки; именам собственным предшествует астериск. Лишь в одном отношении мы отступим от настоящих конвенций: в угоду читабельности примеров для обозначения буквы «ять» вместо знака + используется соответствующий символ Unicode.

тических свойств самой словоформы (падежа, числа, времени, лица, наклонения). В основе автоматической разметки лежит формальная модель словоизменения церковнославянского языка, включающая в себя два основных компонента [18, с. 250–251]:

1. грамматический словарь — перечень лексем с приписанными им словоизменительными параметрами. Это, как минимум, (1) словарная форма и её варианты (при наличии), (2) постоянные признаки лексемы, (3) код парадигмы и частные особенности словоизменения, (4) краткое толкование (по необходимости);
2. грамматическая модель — совокупность таблиц словоизменительных типов (парадигм), в которых задаются системные соотношения между множествами грамматических значений и соответствующих им форм, обозначенные при помощи специальных кодов (индексов).

Обе составляющие модели словоизменения не задаются априорно на основе существующих грамматик и словарей, но эмпирически и итеративно выводятся из содержимого самого корпуса. Таким образом, разработка словаря и модели ведётся параллельно, и взаимные наработки постоянно корректируются и согласуются между собой: с одной стороны, из корпуса постепенно извлекаются ранее не описанные слова, которым при занесении в словарь вручную приписываются леммы и коды парадигматических шаблонов; с другой стороны, по мере обнаружения ранее неучтённых словоизменительных явлений в анализируемом лексическом материале обновляется номенклатура парадигм, пополняется состав грамматических признаков, правил морфонологических чередований и т. д. Суммарный объём словника по состоянию на 2015 г. насчитывал около 35 тыс. лемм и 60 тыс. отдельных словоформ [13, с. 129–131].

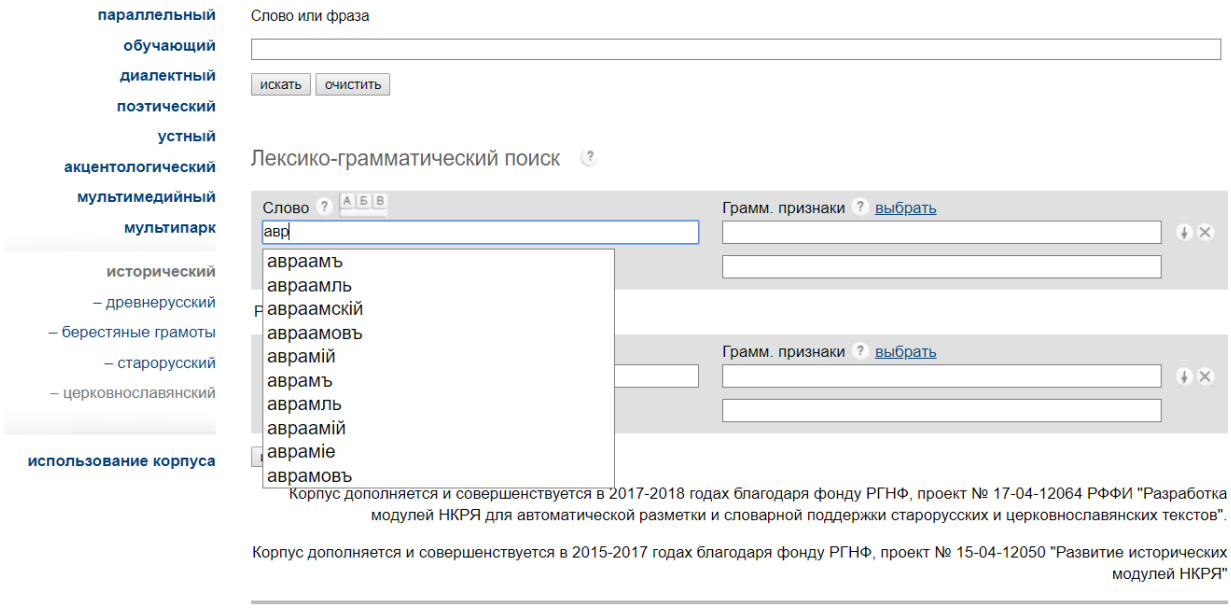
Структура грамматической модели проиллюстрирована в таблице 1.1 (воспроизводится по [18, с. 252]). В заголовочной строке указаны коды па-

Парадигма	**N1t**	**N1t***	**N1j**	**N1k**
Пример	рабъ	сонъ	конь	отрокъ
Основа	раб+ъ	со*н+ъ	кон+ь	отро(к|ц|ч)+ъ
sg,nom	ъ	2ъ	ь	ъ
sg,gen	а	а	я	а
sg,voc	е	е	ю	3е
pl,acc	ы/=gen	ы/=gen	и/=gen	ы/=gen
pl,loc	ѣхъ	ѣхъ	ехъ	2ѣхъ
du,dat/ins	ома	ома	ема	ома

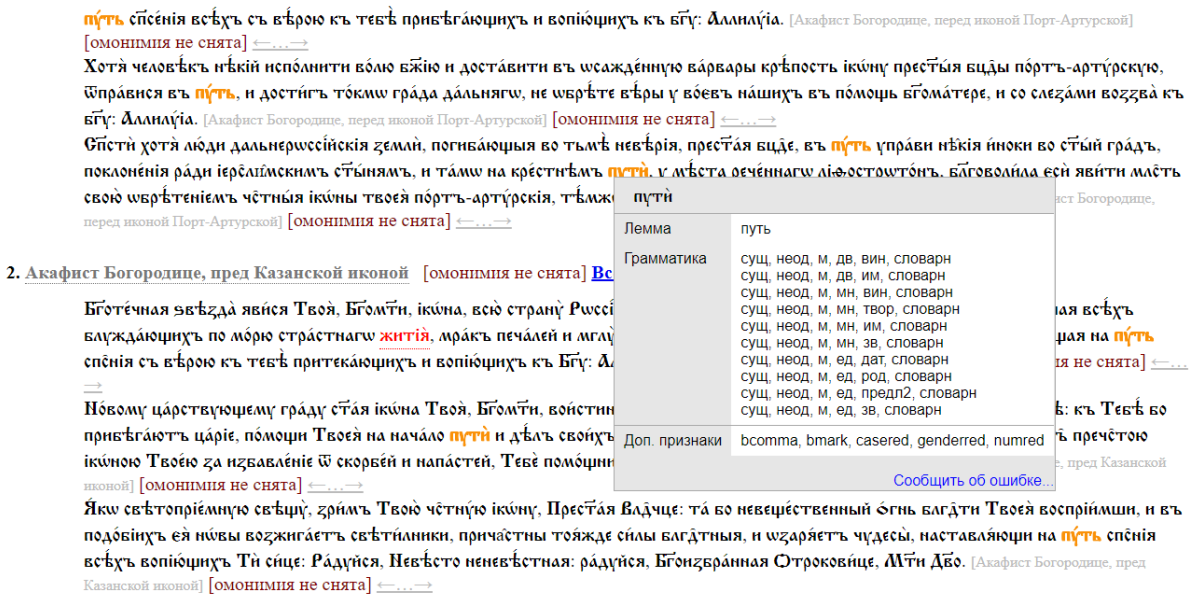
Таблица 1.1. Фрагмент формальной записи парадигм в церковнославянском корпусе

радигм, в двух последующих — типовые примеры лемм в обыкновенной записи и в формате грамматического словаря; далее наборам грамматических признаков в первом столбце сопоставлены наборы соответствующих флексий. Вариативность словарных основ отражается при помощи специальных помет (астериск обозначает беглость предшествующего гласного, в скобки заключаются морфонологические альтернанты), а целочисленные префиксы при флексиях обозначают порядковый номер сочетающегося варианта основы (основным считается первый — без чередований). Всего к настоящему времени составлено не менее 43 таблиц для существительных, 8 для прилагательных, 9 для местоимений-прилагательных, 7 для местоимений-существительных, 50 для глаголов [13, с. 133–134].

Также в цитируемой статье за 2015 г. обсуждается непосредственно эмпирическая реализация модели словоизменения — частотный грамматический словарь церковнославянского языка. Каждой вокабуле частотного словаря (т. е. лемме) помимо словоизменительных параметров, перечисленных выше, сопоставлено суммарное число её употреблений во всех текстах корпуса, за которым следует перечень конкретных словоформ — членов её словоиз-



(*а*) Выпадающий список лемм в поисковой форме



(*б*) Неразрешённая морфологическая неоднозначность

Рис. 1.2. Церковнославянский подкорпус НКРЯ

менительной парадигмы, встретившихся в корпусе (также снабжённых абсолютными частотами встречаемости). От каждой леммы и словоформы можно перейти к цитатам, в которых они содержатся [13, с. 131]. Следует отметить, что страница, запрашиваемая по приводимой А. Е. Поляковым ссылке[10], уже в течение весьма продолжительного времени выдаёт лишь ошибку 404; с другой стороны, при вводе запросов в поисковую форму на основном сайте выпадающий список лемм работает исправно (рис. 1.2, *а*). Вероятно, это свидетельствует о том, что на сегодняшний день частотный словарь уже интегрирован в основной корпус.

Морфологическая дизамбигуация в церковнославянском подкорпусе НКРЯ не производится (рис. 1.2, *б*).

1.1.3. Старорусский корпус

В количественном отношении старорусский (среднерусский) подкорпус НКРЯ превосходит все вышеперечисленные корпуса, вместе взятые: его составляют около 5 тыс. документов XV–XVII вв. суммарным объёмом более 7 млн словоупотреблений. Морфологическая разметка этих текстов также предусмотрена в рамках общей программы развития исторического сегмента НКРЯ, однако, как отмечает Д. В. Сичинава, «её выполнение как вручную, так и автоматически наталкивается на известные сложности, связанные с огромным объёмом текстов и их орфографической и языковой пестротой» [20, с. 227]. Поэтому публичного доступа к грамматической информации данный корпус не предоставляет: организация поиска возможна только по точным совпадениям.

В настоящее время разработки, направленные на изменение такого положения вещей путём автоматизации лексико-грамматической разметки ста-

[10] http://feb-web.ru/febupd/slavonic/dicgram/ (дата обр. 01.06.2018)

рорусского подкорпуса НКРЯ, активно ведутся сотрудниками школы лингвистики НИУ «Высшая школа экономики» под руководством О. Н. Ляшевской. В статье [10] описывается опыт построения двух альтернативных модулей морфологического анализа тестовой выборки из среднерусского корпуса объёмом порядка 2 млн словоупотреблений — словарно-правилового и гибридного.

Первый подход, которому и посвящена основная часть публикации, основывается на церковнославянском грамматическом словаре, обсуждавшемся выше (см. 1.1.2): ввиду отсутствия подобных ресурсов, ориентированных на древнерусский языковой материал (притом сравнительно поздней редакции), авторами делается «сильное допущение, что деление лексики на словоизменительные типы в древнерусском языке в достаточной мере соотносится с делением лексики в церковнославянском» [10, с. 14]. При этом из системы А. Е. Полякова заимствуется исключительно словарная составляющая: в качестве алгоритмической основы морфологического анализатора используется программа «Юни-парсер» Т. А. Архангельского, предназначенная для разметки текстов на языках различных структурных типов. Таким образом, церковнославянский грамматический словарь потребовалось специально адаптировать к механизму работы Юни-парсера — путём решения, в частности, следующих задач [10, с. 14–18]:

1. составление правил для порождения косвенных (т. е. альтернирующих) основ для каждого релевантного словоизменительного типа;
2. введение новых типов склонения, возникших в результате перестройки системы именного словоизменения раннедревнерусского периода;
3. автоматическое порождение основ и предсказание парадигм для глаголов, обладающих неполной грамматической аннотацией в словаре;
4. обработка глаголов, изменяющихся по изолированным типам спряжения либо обладающих нерегулярными особенностями в парадигме.

Второй предлагаемый подход к лексико-грамматической аннотации старорусского корпуса, напротив, не связан с привлечением каких-либо грамматических словарей и опирается исключительно на уже существующие ресурсы. В качестве его методологической основы выступает простое предположение, что, с одной стороны, разбор среднерусских словоформ, имеющих непосредственные современные аналоги, можно без значительного качественного ущерба доверить морфологическим анализаторам для современного русского языка; с другой стороны, прецедентная совокупность ручных разборов, произведённых в рамках древнерусского подкорпуса НКРЯ (см. 1.1.1), может в значительной мере покрыть «архаичные» словоформы (включая исторические формы таких частотных лемм, как глагол быти или местоимение иже), с которыми анализаторы для современного русского языка заведомо не справятся. Для анализа условно-современных словоформ были использованы парсеры MyStem и TreeTagger, адаптированные для НКРЯ [10, с. 18–19].

Заметим, что перед обоими подходами не ставилась дополнительная задача разрешения неоднозначности, а успех каждого конкретного разбора определялся на основании критерия «широкого охвата»: если хотя бы один разбор являлся корректным, всё множество порождаемых разборов также принималось за правильное. С учётом данного обстоятельства приводимые авторами оценки качества как словарного, так и гибридного модуля являются весьма оптимистичными: на золотом стандарте, состоящем из двух вручную размеченных текстов XVI–XVII вв. («Жития Сергия Радонежского» и «Наказа Афанасию Филипповичу Пашкову на воеводство в Даурской земле»), оба парсера показали точность не ниже 93 % и 89 % в случае частеречной разметки и лемматизации соответственно. Показатели полноты и аккуратности словарного парсера ожидаемо ниже, чем у гибридного, однако по точности он неизменно выигрывает — в особенности применительно к более архаичному тексту «Жития» [10, с. 19–21].

Таким образом, не исключено, что в ближайшие годы старорусский корпус, подобно остальным историческим подкорпусам НКРЯ, будет также обогащён собственным лексико-грамматическим инструментарием.

1.2. Регенсбургский диахронический корпус русского языка

Регенсбургский диахронический корпус русского языка (Regensburg Russian Diachronic Corpus, далее RRuDi) во многих отношениях стоит особняком от отечественных восточнославянских корпусов. Его разработка велась при поддержке гранта Немецкого научно-исследовательского общества в рамках проекта «Corpus Linguistics and Diachronic Syntax: The Grammaticalization of Non-Canonical Subjects in Slavonic Languages»; таким образом, RRuDi с самого начала проектировался как эмпирическая база для конкретных исследовательских задач и позиционировался как прикладной ресурс, не стремящийся охватить восточнославянские письменные источники какого-либо временного пласта во всей полноте, — как *диахронический* корпус в противовес *историческим*. Как следствие, документальная основа RRuDi обнаруживает весьма значительное хронологическое и типологическое разнообразие, включая в себя хроники, деловые документы, письма, путевые записки, жития и тексты иных литературных жанров, созданные на протяжении X–XVIII вв. [30, p. 36–37]. Однако в настоящее время для открытого пользования доступен лишь подкорпус общим объёмом порядка 115 тыс. словоупотреблений — для работы с корпусом в полном объёме необходимо заполнить и отправить лицензионное соглашение.

Технологической базой для грамматической аннотации RRuDi является платформа GATE. Разметка производится полуавтоматически — в том смысле, что первичный морфологический анализ осуществляется полностью автоматизированными средствами, но впоследствии его результаты корректи-

руются вручную или путём задания специальных правил [30, p. 43]. Здесь необходимо оговорить то немаловажное обстоятельство, что с точки зрения непосредственного материального источника тексты в составе RRuDi подразделяются на две неравнозначные группы: (1) специально оцифрованные коллективом в процессе проектирования корпуса; (2) изначально опубликованные в серии «Библиотека литературы Древней Руси». На этапе первичного морфологического анализа данные группы обрабатываются по-разному.

Оригинальные тексты RRuDi размечаются полностью автоматически при помощи последовательного применения трёх морфологических анализаторов: старославянского и древнерусского, разработанных в Регенсбургском университете и представляющих собой конечные автоматы типа Xerox, а также статистического разметчика TreeTagger в версии С. А. Шарова [30, p. 44]. Парсеры для обоих древнеписьменных языков создавались независимо от основного корпуса и могут работать автономно: так, старославянский морфологический анализатор был интегрирован в отдельное веб-приложение[11], способное производить грамматический разбор и строить парадигмы по пользовательским запросам.

Механизм обработки несобственных текстов RRuDi несколько менее прямолинеен [29, p. 270–278]. Поскольку в изданиях серии «Библиотека литературы Древней Руси» наряду с самими древнерусскими текстами содержатся их переводы на современный русский язык, разработчиками было принято решение рассматривать данные текстовые массивы как параллельные корпуса и при их морфологическом анализе дополнительно задействовать широко известные в соответствующем домене корпусной лингвистики методы проецирования (annotation projection). Вначале тексты на современном русском и на древнерусском подвергаются процедуре автоматического выравнивания по абзацам и предложениям, а далее подвергаются обработке при помощи

[11] http://rhssl1.uni-regensburg.de:8080/OCS (дата обр. 01.06.2018)

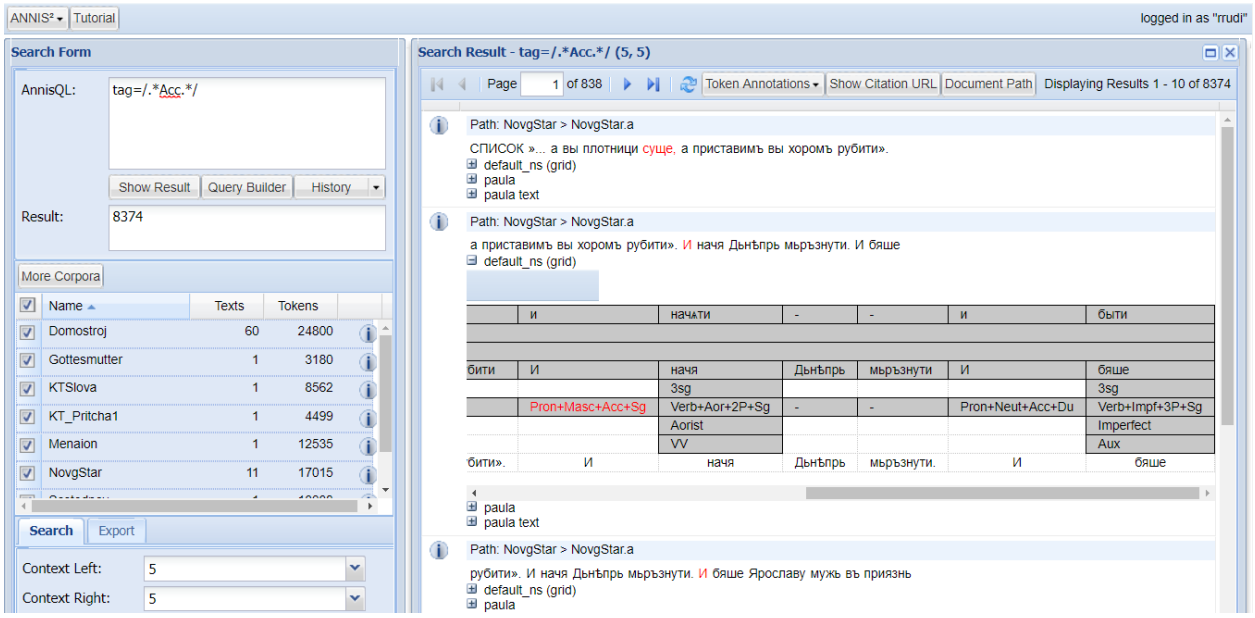


Рис. 1.3. RRuDi: запрос на словоформы в винительном падеже

парсеров, упомянутых выше, — TreeTagger и древнерусского соответственно. Ввиду того что последний функционирует исключительно на основе правил и не основывается ни на каких словарных ресурсах, множества порождаемых им неверных разборов бывают весьма многочисленными; так, Meyer [29, p. 272] приводит следующие примеры того, как им анализируется словоформа НАЧАТИ: (1) инфинитив, (2) краткое страдательное причастие прош. вр. в форме им. п. мн. ч. м. р., (3) существительное ж. р. в форме дат. п. ед. ч., (4) оно же в форме мест. п. ед. ч., (5) оно же в форме им. п. дв. ч.

После промежуточного этапа выравнивания по словам анализатор стремится уменьшить число полученных таким образом древнерусских разборов, проецируя на них те разборы, которые были ранее присвоены выявленным на предыдущем шаге современным русским эквивалентам. При этом он руководствуется рядом специфических правил, например [29, p. 273–274]:

1. если разбор анализируемой древнерусской словоформы единственен, то её разметка остаётся без изменений;
2. если множество древнерусских разборов пусто, на подобную словоформу переносятся все разборы современного русского эквивалента;

3. если у части древнерусских и современных русских разборов совпадают частеречные теги, то все прочие разборы удаляются;
4. если множество современных русских разборов является подмножеством древнерусских, то иные разборы среди последних удаляются.

К сожалению, перечень принятых в RRuDi грамматических тегов нигде специально не приводится, что затрудняет организацию сложных поисковых запросов в корпус-менеджере. Снятие грамматической омонимии не предусмотрено (рис. 1.3).

1.3. Манускрипт

Информационно-аналитическая система (ИАС) «Манускрипт», объединяющая в рамках своих основных коллекций более 140 старославянских (включая 5 глаголических) и древнерусских текстов объёмом более 3,5 млн словоупотреблений, для автоматизированного морфологического анализа последних опирается на электронный грамматический словарь древнерусского языка (ГСДЯ). Формально он представляет из себя «базу данных, содержащую лингвистические единицы, их значения и связи» [8]; среди базовых лингвистических единиц в свою очередь выделяются следующие основные объекты: основа, окончание, тип изменения, вариант основы, парадигма и субпарадигма.

С точки зрения ER-модели (entity–relationship model) ГСДЯ, воспроизведённой на рис. 1.4 по [9], центральное место среди перечисленных компонентов занимают типы изменения. Они представляют собой полноценные словарные единицы, фигурирующие в словнике как индексированные аббревиатуры типа *2а_г*, *а1_прич2* и т. п., и обладают собственными лексико-грамматическими характеристиками; с другой стороны, входящие в них окончания имеют уже непосредственно морфологические параметры (см. таблицу 1.2).

Часть речи	Параметры типов изменения	Параметры окончаний
сущ	род	число, падеж
прил	членность, разряд	род, число, падеж
гл	наклонение	изменяемость, время, число, лицо
прич	*гл* + время, залог, членность	*гл* + род, падеж

Таблица 1.2. Параметры типов изменения и окончаний в ИАС «Манускрипт»

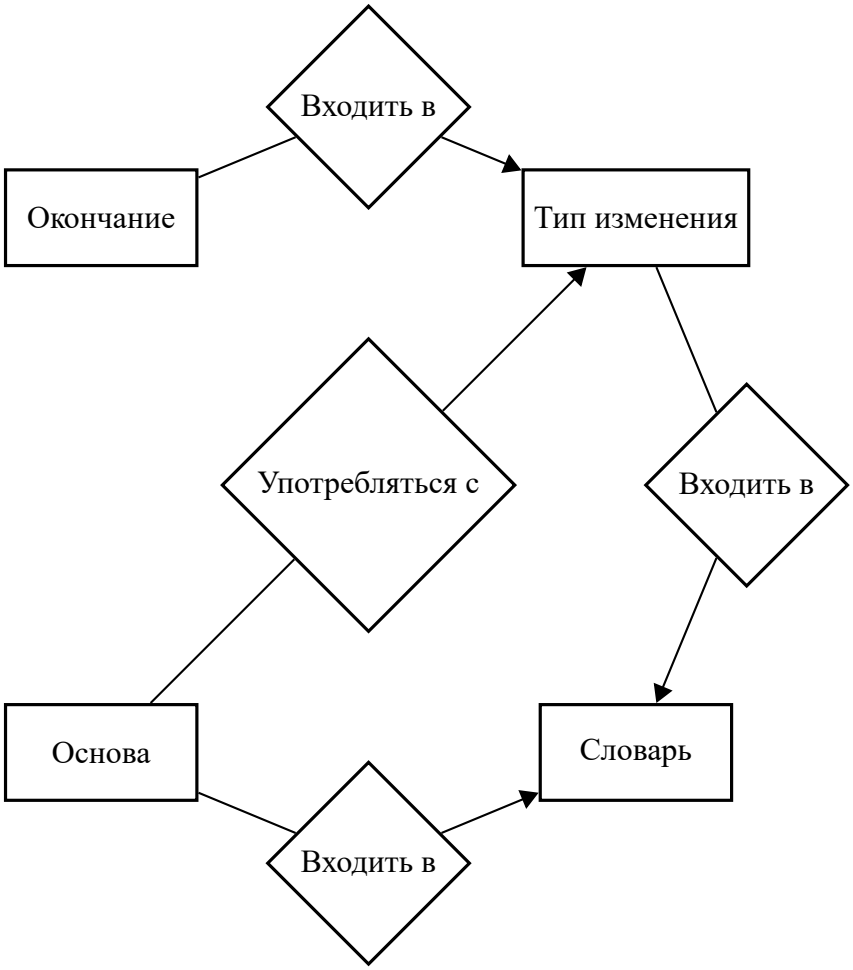


Рис. 1.4. ER-модель ГСДЯ

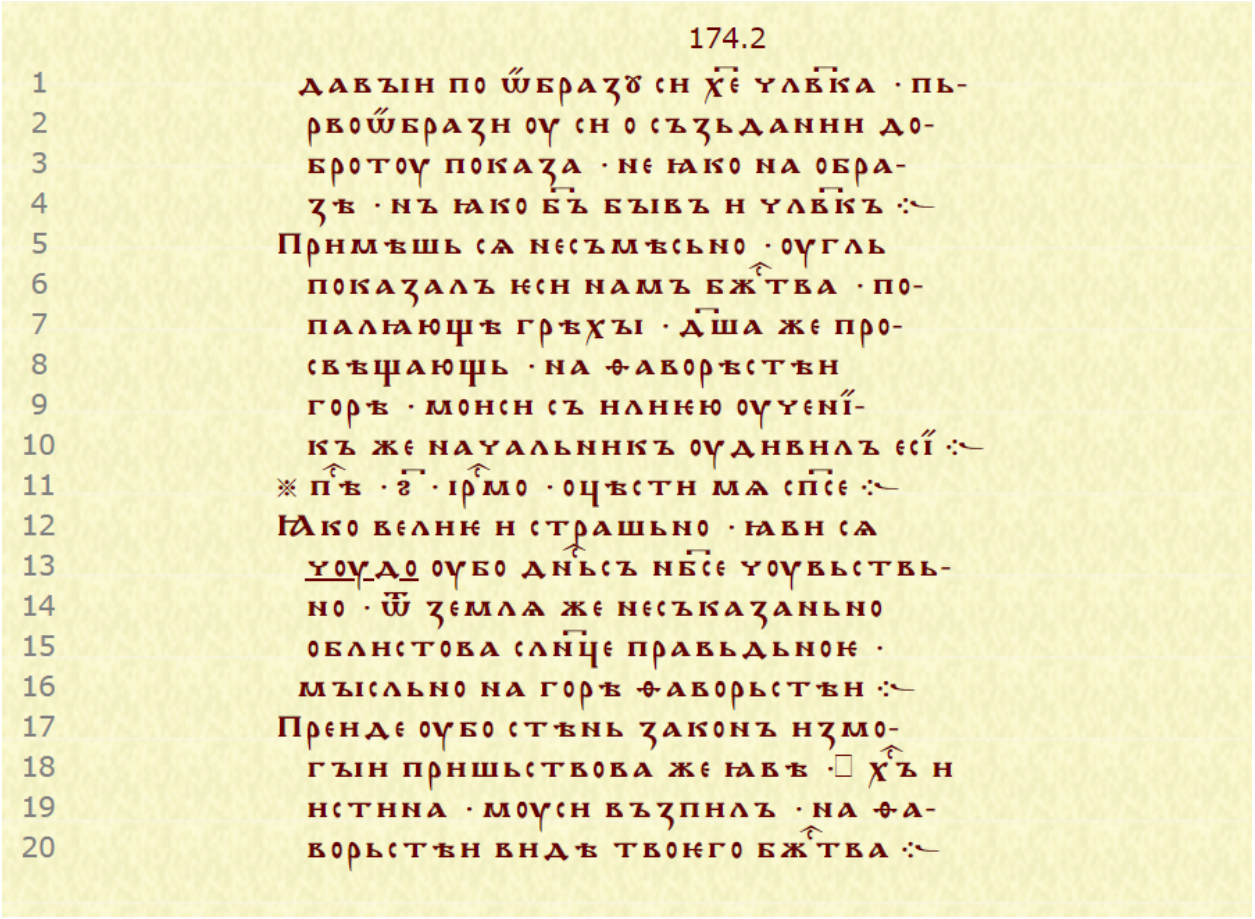
Основы описываются с учётом типа изменения (а также номера омонима и различных лексико-семантических свойств: одушевлённости, собирательности и т. д.), а построение парадигм осуществляется путём конкатенации основ и окончаний одинаковых типов. Выделение в качестве особых словарных единиц вариантов основ и субпарадигм направлено на учёт алломорфирования (ср. -куп- // -купл-) и варьирования типов изменения (например, по признаку членности).

Имплементация морфологического анализатора представлена на сайте ИАС «Манускрипт» (раздел «Инструменты») в четырёх версиях[12], отличающихся друг от друга функциональными возможностями. Первая версия умеет (1) приводить словоформы к начальной форме, (2) выводить грамматические признаки основ и окончаний, (3) строить полные словоизменительные парадигмы, (4) осуществлять поиск словоформ по маске. При этом запросы должны вводиться в поисковую форму исключительно в нормализованном виде: именно такое представление словоформы имеют в базе данных ГСДЯ, а поисковая машина первой версии морфологического анализатора работает только в режиме полного совпадения.

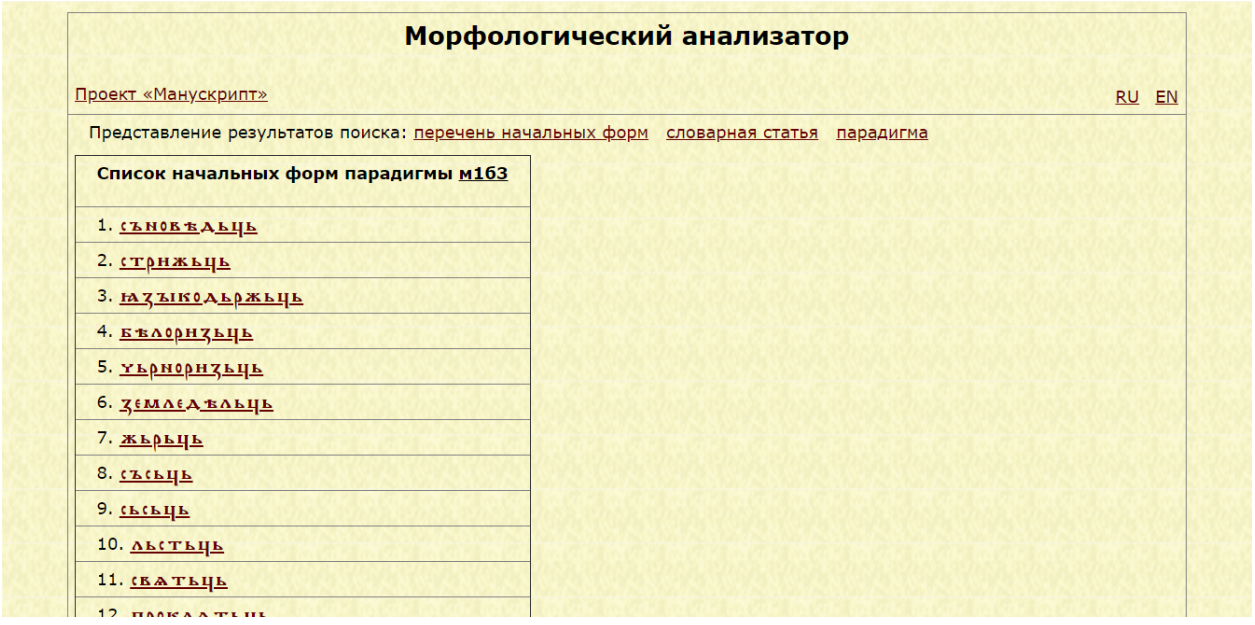
Вторая версия является шагом вперёд по сравнению со своей предшественницей, позволяя приводить к лемме также и графико-орфографические варианты: с одной стороны, для ввода в поисковую форму доступны не только современные кириллические символы, но и исторические; с другой стороны, наряду с поисковыми запросами пользователю предоставляется возможность задавать параметры стандартных орфографических преобразований их внешнего вида. Кроме того, предусмотрено ограничение выводимых результатов необходимыми грамматическими признаками.

Основное новшество третьей версии заключается в возможности морфологического анализа целых текстовых фрагментов — причём как уже раз-

[12] Доступ к пятой версии имеют только создатели транскрипций.



(*а*) Поиск по коллекции славянских миней: выделение запроса чудо



(*б*) Список начальных форм типа изменения *м1б3*

Рис. 1.5. ИАС «Манускрипт»

делённых на словоформы, так и нет (т. н. scriptum continuum — подавляющее большинство текстов из коллекций ИАС «Манускрипт» именно таково). В последнем случае словоделение производится автоматически на основе расчленения поисковых запросов на компоненты, имеющие соответствия в ГСДЯ; ненадёжные варианты устраняются из выдачи. Также в третьей версии реализована локальная дизамбигуация именных групп: при анализе последовательно идущих друг за другом существительных либо местоимений и прилагательных либо причастий, у которых множества разборов пересекаются, в выдачу попадают лишь результаты подобного пересечения.

Наконец, четвёртая версия морфологического анализатора призвана синтезировать наработки, реализованные в двух предыдущих: она позволяет задавать множественные запросы с уточнением расстояния между терминами, а также их грамматических и иных параметров. Однако её наиболее значимое отличие от более ранних версий состоит в возможности искать не только по ГСДЯ, но и по самим текстовым коллекциям: от результатов поиска доступен непосредственный переход к соответствующим рукописным фрагментам, где каждое искомое вхождение выделено подчёркиванием (рис. 1.5, *а*). Не менее важной представляется и возможность просмотра полных перечней основ, имеющих те же типы изменения, что и термины запроса (рис. 1.5, *б*).

1.4. Санкт-Петербургский корпус агиографических текстов

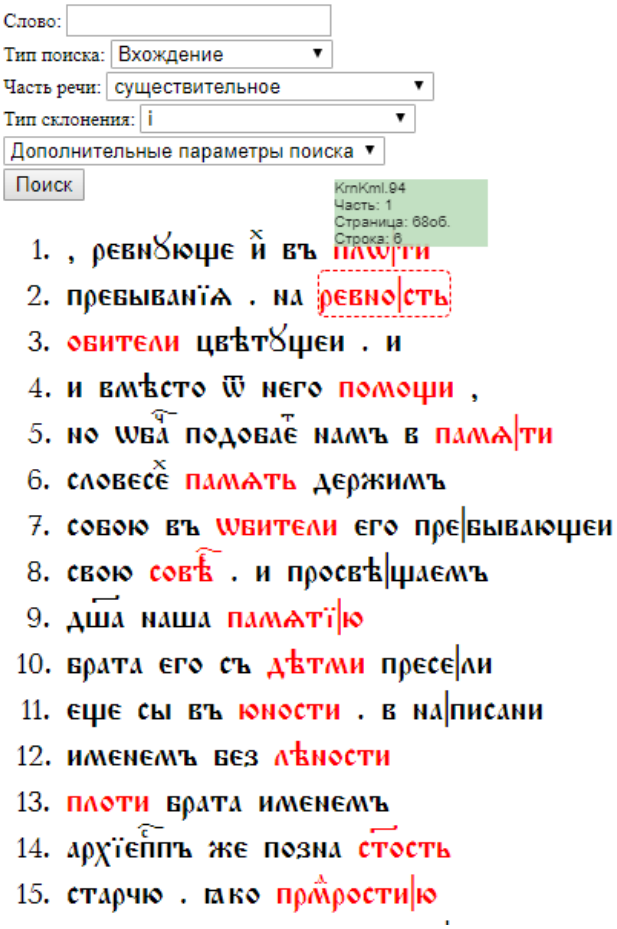
Принятый в СКАТ формат грамматической аннотации был разработан выпускницей кафедры математической лингвистики СПбГУ Е. С. Ивановой [14] и впоследствии видоизменён и уточнён Е. Л. Алексеевой. Он используется для ручного ввода грамматических данных (в течение последнего десятилетия разметка производилась студентами 1–2 курсов в ходе филологической практики) и представляет собой таблицы, где каждой словоформе приписаны

соответствующие ей морфологические (в случае аналитических глагольных форм — также и некоторые синтаксические) характеристики; лемматизация в ходе разметки не осуществляется. К настоящему времени таким образом размечено 5 житий общим объёмом более 50 тыс. словоупотреблений.

Всего для внесения грамматических сведений предусмотрено 6 столбцов, однако фактическое число и значение заполняемых позиций варьирует в зависимости от первой характеристики — части речи. Так, слова знаменательных именных частей речи (существительные, прилагательные и числительные), а также неличные местоимения размечаются единообразно: для них последовательно указываются тип склонения, падеж, число и род; наполнение глагольных тегсетов зависит от наклонения, в случае изъявительного — ещё и от морфологического типа использованного времени (простого или сложного).

Кроме того, формат разметки призван учесть то обстоятельство, что представленные в корпусе житийные тексты, будучи написанными на церковнославянском языке достаточно поздней редакции, отражают живые процессы развития архаичных черт старославянской грамматики: смешение типов склонения, становление категории одушевлённости, обособление деепричастий в самостоятельную глагольную форму и т. д. Для фиксации переходных явлений подобного рода в соответствующей позиции тегсета приводятся два категориальных значения, разделённые косой чертой, — парадигматически ожидаемое и реально встретившееся [5, с. 70–71]. Например:

- тип склонения *es/o* у существительного тѣла обозначает, что исторически его основа относится к одному из подтипов на согласный (*es), но употреблённая флексия соответствует типу *ŏ;
- падеж *вин/род* у существительного бга# показывает, что в значении винительного падежа здесь использован родительный, в чём проявляется категория одушевлённости;



(*а*) Выдача по запросу на существительные *ĭ-склонения

≤ Житие Корнилия Комельского стр. 70 об. ≥

свою совѣ . и просвѣ
щаемъ дша наша памятї
ю wща нашего . w рожс стго .
ѿ славнаго града ростова .
ѿ благочтиваго корене ,
израсте доброплдна розга .
ѿ славны родителеи . иже
мнѡгимъ богатствомъ цвѣ
тущеи , паче всѣ въ гра
дѣ ростовѣ . съ стыи
чдны отрокъ родисѧ .
ѿ оца федора . и мтре
варвары . тъ же федо
не незнаемъ сы и самодер
жавномꙋ всеѧ роꙋсїи .
еже и преселитисѧ

памятї|ю	существительное	i	ж	тв	ед
wца	существительное	jo	м	род	ед
нашего	местоимение	местоим. мягкое	м	род	ед
w	предлог				
рожс	существительное	jo	ср	мест	ед
стго	прилагательное	местоим. твёрдое	м	род	ед
ѿ	предлог				
славнаго	прилагательное	местоим. твёрдое	м	род	ед
града	существительное	o	м	род	ед
ростова (имя собств.)	существительное	o	м	род	ед
ѿ	предлог				
благочтиваго	прилагательное	местоим. твёрдое	м	род	ед
корене	существительное	en	м	род	ед
израсте	глагол	ед	изъяв	аорист простой	3
доброплдна	прилагательное	местоим. твёрдое	ж	им	ед
розга	существительное	a	ж	им	ед
ѿ	предлог				
славны	прилагательное	местоим. твёрдое	м	род	мн

(*б*) Просмотр полного контекста вхождения ПАМЯТІЮ (КК 70 об./2)

Рис. 1.6. Система В. А. Алексеева

- род *ж/м* у причастия БЛГОДАРR# свидетельствует об употреблении формы мужского рода вместо женского — так отражается процесс образования деепричастий.

В приложении А приведён иллюстративный фрагмент морфологической аннотации жития Димитрия Прилуцкого.

В настоящее время массив размеченных текстов хранится сугубо автономно от прочих компонентов лингвистического обеспечения СКАТ, представляя собой материалы, по-прежнему ожидающие интеграции в основной корпус. Задаче исправления подобной ситуации была посвящена практическая часть магистерской диссертации В. А. Алексеева [4, с. 54–64]: в частности, им был разработан механизм внедрения грамматической информации в структуру XML-представления текстов корпуса, а также тестовый вариант полноценной среды для работы с корпусом через интернет, обеспечивающей полнофункциональный поиск как по структурным частям рукописей, так и по грамматическим признакам.

К сожалению, в полном объёме система реализована не была: её онлайн-версия[13] предоставляет доступ лишь к небольшому размеченному фрагменту жития Корнилия Комельского объёмом порядка 10 листов (В. А. Алексеев отмечает, что технические ограничения использованной для размещения системы интернет-площадки не позволяют производить такие ресурсоёмкие операции, как загрузка новых рукописей, автоматически [4, с. 54]). Пример работы с данной средой приведён на рис. 1.6.

Выводы

В таблице 1.3 представлено сопоставление всех рассмотренных в данной главе корпусов по основным релевантным для нас основаниям.

[13] http://scat.v-alexeev.ru (дата обр. 01.06.2018)

Корпус	Период, вв.	Объём, с/у	Разметка	Дизамби-гуация	Леммати-зация
НКРЯ: др.-р.	XI–XIV	500 тыс.	Ручная	Полная	Есть
НКРЯ: б. гр.	XI–XV	20 тыс.	Ручная	Полная	Есть
НКРЯ: ц.-сл.	XVII–XX	4,7 млн	Словарная	Нет	Есть
НКРЯ: ст.-р.	XV–XVII	7 млн	Нет	Нет	Нет
RRuDi	X–XVIII	115+ тыс.	Гибридная	Нет	Есть
Манускрипт	X–XIV	3,5 млн	Словарная	Локальная	Есть
СКАТ	XV–XVII	500 тыс.	Ручная	Полная	Нет

Таблица 1.3. Восточнославянские исторические корпуса: резюме

Нетрудно заметить, что грамматическая аннотация всех перечисленных корпусов (исключая неразмеченный старорусский подкорпус НКРЯ, а также особый во многих отношениях корпус RRuDi) либо опирается на грамматический словарь, либо производится вручную, причём ручная разметка закономерно является необходимым условием для полной дизамбигуации. С другой стороны, лемматизация в каком-либо виде отсутствует лишь в корпусе СКАТ, что подтверждает актуальность поставленной перед настоящей работой цели и крайнюю важность её достижения для приведения СКАТ в соответствие с глобальным уровнем развития восточнославянских исторических корпусов.

Глава 2

Лемматизация церковнославянских словоформ на основе морфологической разметки

В данной главе поэтапно описывается алгоритм лемматизации морфологически размеченных текстов СКАТ, разработанный в ходе практической части настоящей работы; изложение проблем церковнославянского именного словоизменения и иных теоретических вопросов в известной мере вторично и прежде всего обусловлено необходимостью их алгоритмического решения. Программный пакет к текущей и следующей главам написан на языке Python (версии 3.6) и находится в открытом доступе на GitHub[14].

В монографии С. А. Коваля лемматизация понимается как «идентификация инвариантов лексических единиц (выражение ЛО [лексикографического описания. — *К. С.*] с точностью до отдельной лексемы)» [16, с. 76]. Иначе говоря, лемматизация — такая аналитическая процедура, которая для любой входной словоформы позволяет определить, к парадигме какой лексемы она принадлежит, и на выходе эксплицировать словарное наименование этой последней — лемму.

Сразу сделаем два замечания касательно границ применимости реализованного алгоритма. Во-первых, он ориентирован только на морфологически аннотированные словоформы: лемматизация без опоры на какие-либо уже имеющиеся ресурсы (будь то прецедентная разметка или грамматический словарь) принципиально не может быть надёжной и лингвистически корректной в 100 % случаев. Во-вторых, перед работой не ставилась задача охвата глагольного словоизменения: алгоритм способен обрабатывать только словоформы именных частей речи — существительные, прилагательные,

[14] https://github.com/vintagentleman/SCAT

числительные и местоимения, — а также неизменяемые слова (лемматизация которых, впрочем, является тривиальной).

Процедурно в ходе алгоритма, реализованного в настоящей работе, ко входным словоформам последовательно применяются следующие преобразования: (1) орфографическая нормализация, (2) стемминг, (3) восстановление леммы; кроме того, на предварительном этапе подготовки морфологической разметки к обработке в неё вносятся определённые коррективы.

2.1. Корректировка формата морфологической разметки

В ходе нашей работы выяснилось, что степень подробности формата морфологической разметки в корпусе СКАТ, описанного в разделе 1.4, не позволяет учесть все словоизменительные особенности имён, на которые необходимо делать поправку для того, чтобы в дальнейшем лемматизация была произведена корректно. Вследствие этого мы сочли необходимым внести в формат аннотации определённые новшества и уточнения, а также обновить существующие разметки житий Димитрия Прилуцкого, Дионисия Глушицкого и Кирилла Новоезерского с их учётом.

Примеры реализации данных нововведений приведены в таблице 2.1.

2.1.1. Спецификации типов склонения

Тип *o/ja* На материале исследованных текстов этот смешанный тип был зафиксирован только у форм существительного БРАТЪ, чья основа во множественном числе представлена корневым алломорфом -БРАТІ-. Исходя из предположения о том, что и другие существительные с таким смешением обнаруживают подобное алломорфирование (ср. рус. *лист — листья*), данный тип предлагается использовать как маркер наличия йотового наращения у основы и необходимости его удаления при лемматизации.

БРАТIR	сущ	o/ja	им	мн	м	
БРАТIАМИ	сущ	o/ja	тв	мн	м	
ХР(С)ТIАНЕ	сущ	o/en	им	мн	м	
BRTЧRHE	сущ	o/en	тв	мн	м	
DСТА	сущ	o	вин	pt	ср	
ПЕРСИ	сущ	i	вин	pt	ж	
РDЦЂ	сущ	a	вин	дв	ж	*
СТРАСЂ	сущ	o	мест	ед	м	*
МНWЗИ	прил	o	им	мн	м	*
ЕЛЛИНСТIИ	прил	тв	им	мн	м	*
DГЛЂ	сущ	o	мест	ед	м	+o
ПОМЫСЛЫ	сущ	o	вин	мн	м	+e
ЗWЛЪ	сущ	o	род	мн	ср	−o
CDДЕБЪ	сущ	a	род	мн	ж	−e
СТАРЦD	сущ	jo	дат	ед	м	
WВЕЦЬ	сущ	ja	род	мн	ж	
ТОНО(К)	прил	o	вин	ед	м	−o
РА(ДО)СТЕНЪ	прил	o	им	ед	м	−e
ХР(С)ТIАНЫ	сущ	o	вин	мн	м	+ин
ТАТАРЫ	сущ	o	вин	мн	м	+ин
ПО(С)	сущ	o	вин	ед	м	+т
ПЂ(С)МИ	сущ	i	тв	мн	ж	+н
РОЖЕ(Н)И	сущ	jo	мест	ед	ср	+и
БРАНIИ	сущ	ja	вин	ед	ж	−и

Таблица 2.1. Примеры морфологической разметки с учётом нововведений

Тип *o/en* Данный тип присваивается формам множественного числа существительных, обозначающих человека по роду деятельности, происхождению или вероисповеданию с суффиксом -ин- в единственном числе. В парадигме множественного числа данный суффикс утрачивается, и тогда тип *ŏ/*en свидетельствует о том, что в ходе лемматизации его необходимо восстановить.

2.1.2. Существительные pluralia tantum

Существительным, употребляющимся только во множественном числе (pluralia tantum), в данной позиции вместо пометы *мн* присваивается специальная помета *pt*. Её наличие далее (см. 2.4.1) позволяет добавлять к словарным основам подобных существительных флексии множественного числа вместо единственного.

2.1.3. Дополнительные пометы

Ввиду того что в составляющих разметку таблицах последний, шестой столбец при аннотации имён не используется (он задействован только для морфологического описания форм глаголов в форме настоящего-будущего времени и причастий, а также при морфосинтаксическом описании компонентов форм аналитических времён), мы воспользовались этим обстоятельством и отвели его под ряд принципиально новых помет, фиксирующих регулярные морфонологические явления на стыке основы и флексии.

Помета * Астериск в последнем столбце свидетельствует о том, что в абсолютном конце основы размечаемой словоформы действует закон второй палатализации — переход заднеязычных согласных к, г, х в мягкие свистящие ц, з, с перед ѣ или и дифтонгического происхождения. У существительных,

а также у кратких прилагательных вторая палатализация имеет место в следующих парадигматических позициях:

1. тип *ā: дат. п. ед. ч., мест. п. ед. ч., им.-вин. п. дв. ч. (например, рdка — рdцѣ);
2. тип *ŏ, м. р.: мест. п. ед. ч., им. п. мн. ч., мест. п. мн. ч. (rзыкъ — rзыцѣ, rзыцы, rзыцѣхъ);
3. тип *ŏ, ср. р.: мест. п. ед. ч., им.-вин. п. дв. ч., мест. п. мн. ч. (вѣко — вѣцѣ, вѣцѣхъ).

У полных прилагательных (имеется в виду твёрдый тип — ввиду исторической твёрдости заднеязычных согласных) вторая палатализация происходит в (1) дат. п. ед. ч. ж. р.; (2) мест. п. ед. ч. м., ж. и ср. р.; (3) им.-вин. п. дв. ч. ж. и ср. р.; (4) им. п. мн. ч. м. р. (благии — блазѣи, блазѣмъ, блазии). Кроме того, в местоименном склонении имеет место особая разновидность второй палатализации с чередованием -ск- // -ст-: члческии# — члчестѣи# и т. д. [2, с. 139].

Фиксировать случаи второй палатализации на конце основ существительных и прилагательных абсолютно необходимо для корректности процедуры лемматизации, поскольку по формальным признакам невозможно отличить сибилянты, возникшие в результате палатализации, от этимологических, ср.: нозѣ — лемма нога, но трапезѣ — трапеза; дуси — духъ, но бѣси — бѣсъ.

Пометы $\pm o$ и $\pm e$ Наличие одной из указанных четырёх помет указывает на то, что в последнем слоге словарной основы по сравнению с основой размеченной словоформы имеет место прояснение (+) либо падение (−) этимологического редуцированного (соответственно ъ или ь).

Указанные процессы обусловливаются тем, что в последнем слоге формообразующих основ сильные и слабые позиции (см. [24, с. 50–51]) могут че-

редоваться: СОНЪ — СНА, ВЕРЕНЪ — ВЕРНА. Подобное чередование возникает только при наличии в словоизменительной парадигме форм с односложными редуцированными флексиями (современными нулевыми), а следовательно, актуально исключительно для именных парадигм. Более того, размечать его целесообразно отнюдь не для всех словоформ, склоняющихся по именному типу: так, среди несоставных количественных числительных и неличных местоимений данному чередованию подвержены лишь лексемы СТО (род. п. мн. ч. СОТЪ) и ВЕСЬ (род. п. мн. ч. ВСѢХЪ), которые легко поддаются словарному заданию.

Таким образом, пометы $\pm o$ и $\pm e$ приписываются лишь существительным и кратким прилагательным. Реальных случаев употребления указанных помет относительно немного (всего 131 случай на около 11 тыс. существительных и прилагательных, вместе взятых); также отметим следующее:

1. высокочастотные существительные с суффиксом -ц- // -ец- всегда обнаруживают процессы прояснения и падения редуцированного ь и потому обрабатываются автоматически;
2. все прилагательные стандартно приводятся к полным формам (см. 2.4.2), где словарные флексии всегда ненулевые, — а потому при них следует фиксировать только наличие прояснения.

Семейство помет $\pm x$ Если x — произвольная последовательность символов, отличная от *o* и *e*, то соответствующая помета служит для восстановления пропущенных (+) или удаления избыточных (−) букв и буквосочетаний на конце основы при лемматизации.

Подобные пропуски и вставки в большинстве своём носят идиосинкратический характер и во многом обусловлены орфографическим контекстом (например, концом строки) и привычками конкретного писца; тем не менее, иногда постановку указанных помет определяют и причины иного рода.

Например, помета *+ин* регулярно приписывается формам множественного числа существительных, закономерно утративших суффикс деятеля -ин-, но склоняющихся не по ожидаемому типу на согласный (такие случаи предусматривает спецификация типа склонения *ŏ/*en, о которой говорилось выше), а по образцу *ŏ-склонения.

2.2. Орфографическая нормализация

2.2.1. Методологические замечания

Общеизвестно, что в рукописную эпоху орфография не была кодифицирована, и написание многих слов могло существенно варьироваться даже в пределах одного текста. Рассмотрим элементы следующего ряда: БЛАЖЕНАГО, БЛАЖЕННА(Г), БЛ(А)ЖЕННАГО, БЛА(Ж)ЕННАГО, БЛА(Ж)ННА(Г), БЛА(Ж)ННАГО, БЛЖЕНА(Г)#, БЛЖЕНАГО#, БЛЖЕН(Н)АГО#, БЛАЖЕННА(Г)# — очевидно, все они представляют собой варианты записи одной и той же словоформы, которые в целях единообразия обработки на высших языковых уровнях (в т. ч. морфологическом) необходимо предварительно унифицировать путём приведения к единой «нормальной форме» — иначе говоря, подвергнуть процедуре орфографической нормализации[15].

Проблема осложняется тем, что выбор подобной инвариантной единицы также бывает множественным (так, нормальная форма членов вышеприведённого ряда может иметь вид БЛАЖЕНАГО или БЛАЖЕННАГО) и всякий раз обусловливается методологическими установками конкретного исследователя или коллектива. В том случае, если обрабатываемые тексты относятся к начальному периоду жизни древнеписьменного языка, то вне зависимости от датировки всякой конкретной рукописи, как правило, восстанавливается

[15] В компьютерной морфологии термины «нормализация» и «лемматизация» нередко употребляются как абсолютные синонимы [16, с. 75]. Во избежание терминологической путаницы в настоящей работе речь о нормализации идёт исключительно в орфографическом смысле.

каноническая ранняя орфография (и тогда предпочтение было бы отдано варианту БЛАЖЕНАГО); в противном случае выбор осуществляется в пользу современной орфографической нормы (БЛАЖЕННАГО).

Е. Г. Уфлянд, в 2004–2008 гг. работавшая над проблемами нормализации в рамках СКАТ, при обосновании того, что подход, ориентированный на современную орфографию, в контексте рукописей XV–XVII вв. является более целесообразным, опирается на принципы, сформулированные в проекте «Словаря языка житий русских святых XVI–XVII вв.» и во введении к «Словарю русского языка XI–XVII вв.» [22, с. 41–42]: поскольку до XVIII в. церковнославянской орфографической нормы не существовало, нет никакой возможности принимать в качестве эталонного способ написания слов на каком-либо историческом временном срезе; обращаться же к нормативному написанию XIX–XX вв., нежели к современному русскому, не только менее практично, но и бессмысленно.

2.2.2. Модуль нормализации Е. Г. Уфлянд

В практической части своей дипломной работы Е. Г. Уфлянд разработала алгоритм автоматического сведения орфографических вариантов словоформ к основному (т. е. к нормальной форме), реализованный на Python 2.7 в качестве функционального ядра программы для уменьшения объёма сводного словоуказателя. Программой последовательно обрабатывается ряд стандартных ситуаций, в которых наблюдается орфографическое варьирование, например [22, с. 46–70]:

- написания под титлом: БОМТР → БОГОМАТЕР, МЧНК → МУЧЕНИК, ХВ → ХРИСТОВ;
- написания с выносными буквами: Б(Д)Ц / Б(ДИ)Ц → БОГОРОДИЦ, ИС(С)В / ИС(О)В → ИИСУСОВ;

- сочетания типа *TorT с метатезой срединных плавных: млъч / мльч → молч, прьст / пръст → перст;
- регулярные окончания с выносными буквами: а(ш) / я(ш) → аше / яше, бы(с) → бысть.

В ходе интеграции функции, написанной Е. Г. Уфлянд, в нашу собственную программу мы сочли необходимым привнести определённые новшества в механизм её работы. Во-первых, непосредственно в исходный код были внедрены некоторые технологические улучшения:

1. словарная составляющая теперь изолирована от алгоритмической и вынесена в отдельный файл;
2. для хранения информации о буквосочетаниях, подлежащих замене, используются структуры данных типа «словарь» вместо пар синхронизированных между собой массивов;
3. замены были переписаны на языке регулярных выражений, что позволило более полно и экономно охватить вариативность плана выражения сводимых орфографических вариантов.

Во-вторых, сам перечень замен регулярно пополнялся по мере обнаружения нового релевантного языкового материала. Так, при замене ннѣ (и окказионального нне) на нынѣ дополнительно учтены префиксальные деривaты донынѣ, о(т)нынѣ, понынѣ; вообще же в перечень замен сокращённых слов были внесены, например, следующие добавления:

- слв → слав, срц → сердц, стл → святител;
- блг(д)т / блго(д)т → благодат, др(в)н / дрвн → деревн;
- кр(с)тл → крестител, мдр(с)т → мудрост, пр(с)нодв → приснодев.

В-третьих, доступ к данным морфологической разметки позволил корректно производить замены неоднозначных сокращений с выносными буквами и под титлом, которые необходимо раскрывать по-разному в зависимости от принадлежности нормализуемой словоформы к тому или иному лексико-

грамматическому классу. Так, сочетания гн (под титлом) и г(с)дн в начальной позиции форм существительных подлежат замене на господин, в случае же прилагательных — на господн; аналогичным образом ч(с)т приводится к написанию чест либо чист. Выносное (г) на конце прилагательных является стандартным сокращением адъективной флексии род. п. ед. ч. м. и ср. р. и раскрывается как го, однако в абсолютном конце словоформ иных частей речи — как гъ: ср. вра(г), *выпря(г).

Отметим, однако, что в рамках настоящей работы задача окончательного решения проблемы орфографической вариативности в СКАТ не ставилась. Как следствие, такие явления, как (1) чередование редуцированных и гласных полного образования в корнях и префиксах (въсхитити — восхитити), (2) наличие дублетов с одиночными либо удвоенными согласными (воистину — воистинну), (3) непоследовательное написание ятя (грѣхъ — грехъ) [23, с. 378] — в модуле нормализации остаются неучтёнными, частично «процеживаясь» через сито дальнейших этапов алгоритма лемматизации и отражаясь на графическом облике соответствующих лемм.

2.3. Стемминг

2.3.1. Описание алгоритма

На втором этапе алгоритма лемматизации нормализованные словоформы подвергаются процедуре стемминга. Идеологически стеммер, разработанный в рамках данной работы, наследует принципы бессловарно-правиловых стеммеров, основанных на методе усечения окончаний (в частности, имеется в виду классический стеммер Портера [33]). Однако в отличие от формальных подходов, опирающихся исключительно на план выражения и фактически выделяющих в анализируемых словоформах псевдоосновы и псев-

дофлексии (неизменяемые начальные и конечные последовательности символов), реализованный нами алгоритм, напротив, исходит из плана содержания: известные из разметки морфологические свойства обрабатываемых словоформ позволяют максимально точно отделять собственно основы от собственно флексий, лингвистически интерпретируемых и несущих полноценное грамматическое значение, и при этом избегать таких типичных ошибок «слепого» стемминга без опоры на семантику, как over- и understemming.

Так, применение данной процедуры к словоформе день призвано отсечь не псевдофлексию -ень от псевдооссновы д-, но *флексию* -ь от *основы* ден-. Преобразование получаемых таким образом основ с целью их отождествления с основами соответствующих лемм (ср. ден- и дн-) происходит уже на следующем, заключительном этапе алгоритма.

В основе разработанного стеммера лежат словоизменительные парадигмы, составленные нами с опорой на авторитетные учебно-научные пособия по старославянскому языку [2; 15; 24] и программно представляющие собой структуры данных типа «словарь», в которых грамматические значения сопоставлены выражающим их финальным сегментам; последние с целью учёта неустранимой орфографической вариативности записаны в виде регулярных выражений. При анализе каждой входной словоформы сперва осуществляется поиск кортежа из её граммем во множестве ключей парадигмы соответствующего класса (см. 2.3.2); при положительном исходе регулярное выражение, являющееся значением по данному ключу, далее сопоставляется с абсолютным концом анализируемой словоформы. Если результат проверки ненулевой, то найденная подстрока отсекается.

Морфологически неизменяемые формы: несклоняемые прилагательные (близкие по значению к наречиям и весьма немногочисленные [15, с. 140–141]: в каждом из размеченных текстов единожды употреблено лишь прилагательное исполнь ‘полный’), инфинитивы и супины (последние на рас-

смотренном материале не встречаются вовсе), наречия, предлоги и послелоги, союзы, частицы, междометия — стеммингу естественным образом не подлежат; в качестве лемм им присваиваются их нормализованные формы.

2.3.2. Классы словоизменительных парадигм

2.3.2.1. Именные парадигмы

При помощи именных парадигм производится стемминг (1) существительных, (2) нечленных (кратких) прилагательных, (3) несоставных количественных числительных, (4) неличных местоимений в именительном и винительном падежах. Они характеризуются наибольшим качественным разнообразием и значительной вариативностью плана выражения флексий, и для их составления были привлечены дополнительные сведения из монографии [12, с. 257–314].

Например, частная именная парадигма *jā-склонения во множественном числе мужского рода имеет следующий вид:

```
('ja', 'им', 'мн', 'м'): '[+АЕИЫЯ]',
('ja', 'род', 'мн', 'м'): '[+ЕИЫ]И|[ЪЬ]',
('ja', 'дат', 'мн', 'м'): '[АЯ]М[ЪЬ‘]',
('ja', 'вин', 'мн', 'м'): '[+АЕИЫЯ]',
('ja', 'тв', 'мн', 'м'): '[АЯ]МИ',
('ja', 'мест', 'мн', 'м'): '[АЯ]Х[ЪЬ‘]',
('ja', 'зв', 'мн', 'м'): '[+АЕИЫЯ]',
```

2.3.2.2. Местоименные парадигмы

Парадигмы местоименного класса используются при анализе неличных местоимений в косвенных падежах (кроме винительного), а также членных (полных) форм прилагательных и порядковых числительных. Собственно

адъективные флексии фонетически и орфографически весьма близки (их стяжённые разновидности — практически идентичны) местоименным, и уже рукописи X–XI вв. обнаруживают результаты их взаимодействия и уподобления [24, с. 170–171]; исходя из этих соображений мы сочли приемлемым объединить их в общий класс.

В качестве примера приведём парадигму местоименного склонения по твёрдому типу в единственном числе среднего рода:

```
('тв', 'им', 'ед', 'ср'): 'О?Е',
('тв', 'род', 'ед', 'ср'): 'А?[АЕО]?ГО',
('тв', 'дат', 'ед', 'ср'): 'У?[ЕОУ]?МУ',
('тв', 'вин', 'ед', 'ср'): 'О?Е',
('тв', 'тв', 'ед', 'ср'): '[ИЫ]?[+ЕИЫ]М[ЪЬ‘]',
('тв', 'мест', 'ед', 'ср'): '[+Е]?[+ЕО]М[ЪЬ‘]',
('тв', 'зв', 'ед', 'ср'): 'О?Е',
```

2.3.2.3. Особые местоименные парадигмы

Словоизменению таких местоимений, как (1) личные азъ, ты (ед. ч.), вѣ, ва (дв. ч.), мы, вы (мн. ч.), (2) возвратное себе, (3) вопросительные кто и что, — присущ ряд глубоко архаических черт (ярко выраженный супплетивизм, особая система флексий), из чего следует принципиальная невозможность их анализа по общим правилам. Их словоизменительные парадигмы обособлены от парадигм прочих классов и имеют несколько иную структурную организацию; описывающие их словоизменение регулярные выражения представляют собой полные покрытия соответствующих символьных цепочек, а обращение к разметке производится исключительно с целью выявления в ней возможных ошибок.

Рассмотрим примеры парадигм местоимений всех перечисленных разрядов:

```
('1', 'им', 'ед'): ('АЗ[ЪЬ‘]?$', 'АЗЪ'),
('1', 'род', 'ед'): ('М([ЕЪЬ]?Н)?[+ЕЯ]$', 'АЗЪ'),
('1', 'дат', 'ед'): ('М([ЕЪЬ]?Н[+Е]|И)$', 'АЗЪ'),
('1', 'вин', 'ед'): ('М([ЕЪЬ]?Н)?[+ЕЯ]$', 'АЗЪ'),
('1', 'тв', 'ед'): ('М[ЪЬ]?НОЮ$', 'АЗЪ'),
('1', 'мест', 'ед'): ('М[ЪЬ]?Н[+Е]$', 'АЗЪ'),

'род': ('С([ЕО]Б)?[+ЕЯ]$', 'СЕБЕ'),
'дат': ('С([ЕО]Б[+Е]|И)$', 'СЕБЕ'),
'вин': ('С([ЕО]Б)?[+ЕЯ]$', 'СЕБЕ'),
'тв': ('СОБОЮ$', 'СЕБЕ'),
'мест': ('С[ЕО]Б[+Е]$', 'СЕБЕ'),

('м', 'им'): ('Ч[ЪЬ]?ТО$', 'ЧТО'),
('м', 'род'): ('Ч[ЕЬ]?(СО)?(ГО)?$', 'ЧТО'),
('м', 'дат'): ('Ч[ЕЬ]?(СО)?МУ$', 'ЧТО'),
('м', 'вин'): ('Ч[ЪЬ]?ТО$', 'ЧТО'),
('м', 'тв'): ('ЧИМ[ЪЬ‘]?$', 'ЧТО'),
('м', 'мест'): ('Ч[ЕЬ]?(СО)?М[ЪЬ‘]?$', 'ЧТО'),
```

Полностью аналогично вопросительным местоимениям обрабатываются неопределённые нѣкто, нѣчто и отрицательные никтоже, ничтоже.

Наконец, особого упоминания здесь заслуживает определительное местоимение кождо ‘каждый’[16], в косвенных падежах изменяющееся по образцу вопросительного кто (при этом сегмент -ждо- ведёт себя подобно слитной частице же и не оказывает влияния на словоизменение): когождо, комуждо [I, 21, с. 1389]. Отметим также, что наряду с ним существует синонимичное местоимение киижде (с вариантом коижде), однако оно, в отличие от кождо, имеет регулярную парадигму на основе вопросительного кии: коегождо, коемуждо [I, 21, с. 1417] — и потому обрабатывается по общему правилу.

[16]В «Материалах» И. И. Срезневского также приводятся примеры с конечным сегментом -жде-, но в корпусе подобных употреблений зафиксировано не было.

2.3.3. Обработка составных форм

2.3.3.1. Составные существительные

На материале рассмотренных житийных текстов выделяются две семантико-морфологические группы составных существительных со склонением обеих частей в их составе.

Во-первых, речь идёт о названиях населённых пунктов со второй корневой морфемой -ГРАД- либо -ГОРОД- (полногласный вариант низкочастотен, но встречается в некоторых текстах корпуса): *КОСТЯНТИНЬГРАДЪ, *НОВЪГРАДЪ[17]. При анализе подобных имён собственных мы исходим из посылки, что их первая составляющая всегда представляет собой краткое прилагательное мужского рода *ŏ- или *jŏ-склонения, т. е. изменяется аналогично существительному ГРАДЪ. Таким образом, тип склонения первой части можно считать известным, а процедуру стемминга единообразно производить над обоими компонентами: *НОВЂГРАДЂ → -НОВ-, -ГРАД-.

Во-вторых, спецификой в рассматриваемом аспекте обладают наименования времён суток ПОЛДЕНЬ и ПОЛНОЩЬ (русизм ПОЛНОЧЬ в корпусе не встречается). Здесь типы склонения составных частей не совпадают: существительные ДЕНЬ и НОЩЬ склоняются по типам *en и *ĭ соответственно, а ПОЛЪ (в значении ‘половина’) — по типу *ŭ. Однако несмотря на то, что последний нам априорно известен, и при соответствующей поправке первая часть также подлежит стеммингу, ввиду предельной ограниченности и закрытости данной группы существительных было решено проверять их начальные подстроки на соответствие простым регулярным выражениям: `ПОЛ.*Д[ЕЬ]?Н` либо `ПОЛ.*НО[ЧЩ]` — и при положительном результате приписывать готовые леммы без какого-либо анализа грамматических данных.

[17] Слитное написание подобных топонимов — результат модернизации орфографии: в самих рукописях XV–XVII вв. они представляют собой словосочетания вида «краткое прилагательное + существительное» с последовательным склонением обеих лексем.

2.3.3.2. Составные числительные

Церковнославянские сложносоставные количественные числительные функционируют как словосочетания, и поэтому при слитном написании им присущи специфические словоизменительные особенности, отчасти присутствующие и в современном русском языке.

1. Числительные, обозначающие числа от 11 до 19, представляют собой сочетания единиц первого десятка с предложно-падежной группой на десяте (мест. п.). Формальный тип синтаксической связи внутри подобных структур — предложно-падежное примыкание; изменению подвержена только первая часть: пятьнадесяте — пятинадесяте.

2. Названия чисел 20, 30, 40 и 200, 300, 400 являются сочетаниями имён соответствующих единиц с существительными десять (склоняется по типу *ent) либо сто (типа *ŏ). Тип связи — согласование, а следовательно, при склонении изменяются оба компонента: двадесяти — двудесяту (дв. ч.), триста — трехъсотъ (мн. ч.).

3. Обозначения чисел от 50 до 90 и от 500 до 900 лексически подобны названиям десятков и сотен меньших порядков, однако синтаксически ведут себя иначе: здесь наименования единиц управляют существительными десять или сто в форме род. п. мн. ч. Изменяется только первая часть: седмьдесятъ — седмидесятъ, осмьсотъ — осмисотъ.

Принимая во внимание закрытость множества составных числительных и стремясь избежать излишних технологических затруднений, которые могли бы возникнуть при их прямолинейном анализе, здесь мы также ограничились описанием морфемной структуры в виде регулярных выражений:

```
'ЕДИН.*НАДЕСЯТ': 'ЕДИННАДЕСЯТЕ',
'Д[ЪЬ]?В.*НАДЕСЯТ': 'ДВАНАДЕСЯТЕ',
'ТР.*НАДЕСЯТ': 'ТРИНАДЕСЯТЕ',
```

```
'ЧЕТЫР.*ДЕСЯТ': 'ЧЕТЫРЕДЕСЯТЕ',
'ПЯТ.*ДЕСЯТ': 'ПЯТЬДЕСЯТЪ',
'ШЕСТ.*ДЕСЯТ': 'ШЕСТЬДЕСЯТЪ',

'СЕДМ.*С[ЪО]?Т': 'СЕДМЬСОТЪ',
'ОСМ.*С[ЪО]?Т': 'ОСМЬСОТЪ',
'ДЕВЯТ.*С[ЪО]?Т': 'ДЕВЯТЬСОТЪ',
```

2.4. Восстановление леммы

Основы, получаемые в результате стемминга, не всегда совпадают с основами соответствующих лемм и потому до прибавления словарных флексий зачастую нуждаются в дополнительных модификациях. Прежде всего это те преобразования, которые продиктованы специальными пометами в самой скорректированной разметке (см. 2.1); однако наряду с ними существуют и такие, которые программа способна производить автоматически.

2.4.1. Существительные

Из существительных следующих подтипов склонения на согласный: *ent, *men, *es, *er — удаляются тематические суффиксы: -врѣмен- → -врѣм-, -словес- → -слов- и т. д. С другой стороны, в случае отсутствия тематических суффиксов в составе основ существительных подтипов *en и *ū последние, напротив, восстанавливаются из соображений модернизации лемм (архаичные формы им. п. ед. ч. без осложнения довольно рано начали замещаться формами вин. п. [15, с. 126]): -кам- → -камен-, -люб- → -любов-. Сходным образом восстановлению подлежит субморф -ос- у частотного имени собственного *христосъ, подвергающийся утрате в косвенных падежах.

Кроме того, основы существительных (и среди имён только их) в отдельных парадигматических позициях подвергаются закону первой палата-

лизации — переходу заднеязычных согласных к, г, х в мягкие шипящие ч, ж, ш в положении перед гласными переднего ряда, а также свистящего ц в составе суффикса деятеля мужского пола и з в слове князь. Фактически сфера действия данного закона ограничена (1) формой зв. п. ед. ч. в парадигме *ŏ-склонения м. р.: бгъ# — бже#, отець — отче (в последнем случае имеет место смешение *jŏ/*ŏ); (2) формами дв. и мн. ч. двух существительных, обозначающих части тела: око — очи (смешение *es/*ĭ) — очеса и аналогично ухо — уши — ушеса. Первая палатализация, в отличие от второй, всегда предсказуема и потому устранима автоматически.

К словарным основам прибавляются флексии им. п. ед. ч. (мн. ч., если данное существительное размечено как plurale tantum) с последовательным учётом типа склонения и родовой принадлежности.

2.4.2. Прилагательные

Из форм прилагательных сравнительной степени (они размечаются при помощи особого частеречного тега *прил/ср*) удаляется суффикс -ш- (включая алломорфы), присущий всем членам словоизменительной парадигмы, кроме им. п. ед. ч. м. и ср. р. Заметим, однако, что мы считаем компаратив самостоятельной грамматической категорией и никаких дальнейших преобразований, нацеленных на перевод сравнительной степени в положительную (в т. ч. устранение супплетивизма типа болшии — великии), не производим.

К словарным основам всех прилагательных прибавляются местоименные флексии им. п. ед. ч. м. р. -ыи (к основам твёрдой разновидности, кроме основ на заднеязычные) или -ии (ко всем прочим): непорочныи, божии; к несупплетивным формам компаратива (на гласную) добавляется -и: грѣшнѣи. Исключение здесь составляют имена собственные: к их основам наряду с местоименными флексиями могут добавляться и именные, а также сохраня-

ется их родовая характеристика: бѣлыхъ (ризахъ) — лемма бѣлыи, но *бѣла (*езера) — *бѣло.

2.4.3. Местоимения

Основы местоимений внутри словоизменительных парадигм практически не обнаруживают вариативности. Точечные модификации требуются в отдельных частных случаях: -ко- → -к- (кии — коего, коему и т. д.), -с- → -се- (унификация альтернантов сеи и сии в пользу более частотного); из предложных форм местоимения и: него, нему и т. д. — удаляется наращение -н- (его основа, таким образом, формально считается нулевой). В остальном построение местоименных лемм осуществляется аналогично прилагательным, и лишь нескольким группам местоимений приписываются специфические конечные сегменты: (1) вашь, весь, нашь, сиць; (2) онъ, самъ; (3) елико; (4) онсица; (5) тои; (6) и.

2.4.4. Числительные

Среди числительных вариативный характер имеет лишь основа -об-: оба — обоихъ. В иных преобразованиях они не нуждаются и сразу дополняются соответствующими именными флексиями им. п.: дв. ч. — числительные два и оба, мн. ч. — три и четыре, ед. ч. — все прочие.

Выводы

Описанный в настоящей главе алгоритм разрабатывался нами методом последовательных приближений: по составлении первичных вариантов словоизменительных парадигм и запуске программы на размеченном материале в выдаче выявлялись необработанные случаи и систематические ошибки, которые далее устранялись либо на уровне разметки, либо алгоритмически.

В результате нам удалось добиться того, что среди 29617 словоупотреблений в составе житий Димитрия Прилуцкого, Дионисия Глушицкого и Кирилла Новоезерского программой обрабатываются 24319 словоформы — 14858 имён и 9461 неизменяемое слово. Экспертная оценка полученных результатов позволяет утверждать, что на материале указанных агиографических текстов значения точности и полноты разработанного алгоритма составляют 100 %: всем анализируемым словоформам присваиваются корректные леммы, необработанных случаев нет.

Глава 3

Загрузка корпуса СКАТ на платформу TXM

Платформа TXM[18] — это свободно распространяемое программное обеспечение для работы с текстовыми корпусами, разработанное в лаборатории IHRIM (Institut d'Histoire des Représentations et des Idées dans les Modernités) Национального центра научных исследований Франции [26]. TXM предоставляет в распоряжение пользователя широкий набор инструментов количественного и качественного анализа текстов: получение конкордансов в формате KWIC и частотных списков лексических единиц на основе любого приписанного им параметра; построение частотных графиков динамики вхождений единиц, удовлетворяющих пользовательскому запросу (для статистических расчётов используется вычислительный движок R); сбор данных о совместной встречаемости, о лексических шаблонах и многое другое. Также платформа приспособлена для обработки текстовой метаинформации, что позволяет пользователю строить подкорпуса (subcorpora) и разбиения (partitions) корпусов, введённых в платформу, по различным метатекстовым основаниям. TXM поддерживает множество входных форматов (TXT, ODT/DOC/RTF, XML, различные проприетарные форматы), однако для внутреннего представления содержимого введённых корпусов используется XML-представление.

По инициативе А. М. Лаврентьева, одного из главных разработчиков платформы, на протяжении нескольких лет активно сотрудничавшего с коллективом СКАТ и впервые написавшего программу для автоматической конвертации текстовых файлов житий в формат XML [6, с. 21], фрагмент корпуса СКАТ объёмом 12 житийных текстов (включая 2 похвальных слова) был за-

[18] http://textometrie.ens-lyon.fr (дата обр. 01.06.2018)

гружен на демонстрационный портал TXM, открытый для пользования в режиме онлайн[19]. Однако сотрудники СКАТ участия в этой работе фактически не принимали, вследствие чего корпус был не вполне качественно адаптирован к реалиям платформы: в частности, сами тексты доступны для чтения лишь в упрощённой графике и содержат ошибки перекодирования (в особенности это касается буквенных обозначений чисел).

Настоящая глава посвящена процессу устранения всех подобных недостатков и максимального приспособления корпуса СКАТ к комфортному использованию при помощи стационарной версии платформы TXM, а также внедрения в TXM-совместимое представление текстов корпуса слоя грамматических данных и лемм.

3.1. Режим импортирования XTZ

Как было отмечено ранее, платформа TXM приспособлена к импорту текстовых корпусов во множестве различных форматов, однако де-факто стандартным и наиболее активно совершенствуемым в позднейших версиях платформы способом загрузки входных текстов в формате XML является режим XTZ — XML TEI Zero [28, p. 76].

Помимо универсальных средств обработки импортируемых документов (включая транспонирование различных уровней разметки во внутреннее TXM-представление, благодаря которому пользователь получает возможность строить подкорпусы и разбиения по любым интересующим его размеченным текстовым структурам, многоаспектное индексирование словоформ и многое другое), режиму XTZ также присуща ориентированность на определённый минимальный («нулевой») набор тегов, наиболее часто используемых при разметке текстовых данных с опорой на рекомендации консорциума

[19] http://portal.textometrie.org/demo/ (дата обр. 01.06.2018)

XML	HTML	Пояснение
<head>	<h2>	Заголовок
<p>	<p>	Абзац
<hi>	<b>	Полужирное начертание
<emph>	<i>	Курсивное начертание
<list type='unordered'>		Маркированный список
<list type='ordered'>		Нумерованный список
<item>		Элемент списка
<table>	<table>	Таблица
<row>	<tr>	Табличная строка
<cell>	<td>	Табличная ячейка
<graphic>	<img>	Рисунок
<ref>	<a>	Гиперссылка
<note>	<a> +	Сноска
<w>		Токен

Таблица 3.1. Преобразования тегов при импорте в режиме XTZ

TEI, и способность учитывать их семантику при конструировании HTML-изданий, непосредственно доступных для чтения.

Так, определяемые TEI маркеры начала новой строки — <lb/> (line beginning) — при генерации HTML преобразуются в теги
, позволяющие форсировать разрыв строки в любом необходимом месте. Кроме того, если они дополнительно снабжены глобальным атрибутом @n, указывающим на порядковый номер соответствующей строки, то напротив строк через определённые интервалы автоматически вставляются их порядковые номера, подобно тому как нумеруются стихи в академических изданиях античной поэзии. Аналогично обрабатывается тег <pb/> (page beginning); в тех случаях, когда пагинация текстов корпуса на уровне разметки не предусмотрена, TXM фрагментирует их самостоятельно, исходя из максимального числа токенов

на каждой странице (этот параметр задаётся пользователем при импорте).

Перечень всех поддерживаемых режимом XTZ XML-тегов и их HTML-эквивалентов приведён в таблице 3.1 (по [28, p. 78–80]).

3.2. Обновление XML-представления СКАТ

3.2.1. Проблемы существующей XML-структуры

Последним, кто работал над СКАТ в рассматриваемом аспекте, был В. А. Алексеев: в рамках своей магистерской диссертации [4, с. 41–54] он предпринял ряд серьёзных мер, направленных на модернизацию XML-представления текстов СКАТ в соответствии с современными стандартами электронного представления текстовых данных.

Нестандартные сущности, теги и атрибуты, ранее использовавшиеся для отображения графем, отсутствующих в современном русском языке, были заменены на символы Unicode 5.1. Данная мера была продиктована как нормативными, так и прагматическими соображениями, поскольку XML-представление СКАТ образца нулевых было весьма громоздким и неудобочитаемым; так, результат преобразования в XML такой словоформы, как РА(Д)УАСR, в нём выглядел следующим образом:

```
pa<osl_letter type='overline'>
  д
</osl_letter>yac&cyr-littleyus;
```

Те немногие графемы, которые не были определены в кодовой таблице Unicode 5.1, В. А. Алексеев предложил по-прежнему кодировать как сущности — например, `&i8-overline;` в случае выносного и восьмеричного. При этом все сущности были определены в отдельном файле определения типа документа (DTD), а также снабжены формальной декларацией (`<charDecl>`)

на уровне описания кодировки TEI-документа (<encodingDesc>). Этот механизм был впервые включён в рекомендации TEI в версии P5 [34, p. 39, 192–201], полноценное обновление до которой и строгое следование соответствующим нормам также входило в круг задач диссертационного исследования.

Тем не менее, разработанная В. А. Алексеевым версия XML-представления СКАТ по ряду причин не является TXM-совместимой. Во-первых, для разметки мельчайших структурных частей рукописи (страниц, колонок и строк) было предложено использовать сразу два синонимичных набора элементов:

1. парные теги <div2>, <div3>[20], <div4>, <l>;
2. пустые теги <pb/>, <cb/>, <lb/>.

Если первый набор нацелен на описание формально-иерархической организации XML-документа, то последний скорее предназначен для его семантического структурирования: вместо разбивки на строго непересекающиеся блоки в текст вносятся маркеры (milestones), попросту указывающие на окончание одной структурной единицы и начало другой. Оба способа одновременно консорциум TEI предписывает задействовать лишь тогда, когда размечаемых структур более одной и они являются соперничающими [34, p. 123–124], т. е. синонимичными, но не идентичными, — однако разбиение текста на такие физические единицы, как строки, колонки и страницы, очевидно, едва ли может обнаруживать какие-либо существенные противоречия. Использовать для них именно маркерную аннотацию предпочтительнее потому, что в противном случае при потенциальном введении в разметку новых структурных слоёв возникает риск их пересечения друг с другом (а запрет на пересекающиеся теги продиктован самым форматом XML). Кроме того, спецификации режима XTZ затрагивают именно пустые теги, а использования их парных аналогов (и шире — всех элементов и атрибутов с целочисленными суффик-

[20] Тег <div2> маркирует лист, а <div3> — страницу, т. е. одну из сторон листа (лицевую либо оборотную).

сами) ввиду особенностей функционирования поисковой машины CQP, напротив, рекомендуется избегать [28, p. 78].

Во-вторых, XML-разметка элементарных лексических единиц (токенов) была призвана учесть множество различных вариантов их графического представления. При этом все подобные варианты определялись как потомки базового тега <w>:

```
<w xml:id='CrlNvz.1'>
  <orig>мѣсѧца</orig>
  <reg>М+СЯЦА</reg>
  <src>М+СRЦА</src>
</w>
```

Здесь внутри вложенного тега <src>[21] первое слово жития Кирилла Новоезерского представлено в оригинальном 8-битном формате, внутри <reg> — в упрощённой графике; наконец, в теге <orig> оно записано с использованием символов Unicode 5.1. В случае ошибочных написаний иерархия получает дальнейшее усложнение: <orig> в качестве потомка приобретает тег <choice>, обозначающий наличие альтернантов[22], а внутрь него в свою очередь заносится ошибка (<sic>) и исправление (<corr>). Например:

```
<w xml:id='CrlNvz.90'>
  <orig><choice>
      <sic>человѣчетвw</sic>
      <corr>человѣчествw</corr>
  </choice></orig>
  <reg>~ЧЕЛОВ+ЧЕТВО &lt;ЧЕЛОВ+ЧЕСТВО&gt;</reg>
  <src>~ЧЕЛОВ+ЧЕТВW &lt;ЧЕЛОВ+ЧЕСТВW&gt;</src>
</w>
```

[21] Это единственный случай отступления В. А. Алексеевым от рекомендаций TEI. Паронимический тег <source> имеет совершенно иную семантику и иное назначение [34, p. 356].

[22] Строго говоря, триада из <orig>, <reg> и <src> также требует обрамления тегом <choice>, однако подобный шаг, очевидно, ознаменовал бы собой ещё большее осложнение XML-структуры.

Между тем режим XTZ не предполагает наличия у ядерных лексических единиц столь развитой иерархической организации. Он ориентирован на обработку тегов `<w>` в простейшем виде, когда в качестве их содержимого выступает единственный вариант графического представления токена, а все альтернативные наряду с прочими сопутствующими сведениями записаны в атрибуты [28, p. 77]. Иначе говоря, предполагается, что элементы `<w>` являются терминальными узлами XML-структуры и потомков не имеют; если в действительности это не так, то при импорте последние игнорируются, а содержимым родительского тега считается результат конкатенации содержимого всех дочерних.

Таким образом, при подготовке HTML-издания первый пример из приведённых выше считался бы тождественным следующему (что было бы нежелательно):

```
<w xml:id='CrlNvz.1'>
  мЂсѧцаМ+СЯЦАМ+СRЦА
</w>
```

3.2.2. Структурные нововведения

Предлагаемые нами нововведения в XML-структуру текстов СКАТ, призванные обеспечить их полную совместимость с режимом импортирования XTZ, обобщены в таблице 3.2.

С чисто формальной точки зрения замещение структурных подразделений верхнего уровня `<div1>` анонимными блоками `<ab>` (anonymous block) обусловлено обозначенным выше стремлением избавиться от тегов с целочисленными суффиксами; содержательная же подоплёка данного нововведения состоит в том, что так житийные тексты представляются как нерасчленённые, — иначе говоря, делается имплицитное утверждение, что никаких промежуточных блоков внутри них не выделяется. Однако в будущем такое

Старый тег	Новый тег
`<div1 type=’part’ n=’1’></div1>`	`<ab></ab>`

`<div2 type=’page’ n=’1’>`
`  <div3 type=’back’></div3>`
`</div2>`
	`<pb n=’-1’/>`

`<div3 type=’front’>`
`  <div4 type=’col’ n=’1’></div4>`
`</div3>`
	`<pb n=’1a’/>`

`<l n=’1’></l>`	`<lb n=’1’/>`

`<w>`
`  <orig>w̄</orig>`
`  <reg>O(T)</reg>`
`  <src>W(T)</src>`
`</w>`
	`<w reg=’o(т)’ src=’W(T)’>w̄</w>`

`<w>`
`  <orig><choice>`
`  <sic>мь</sic>`
`  <corr>мʌ</corr>`
`  </choice></orig>`
`  <reg>~Mь <МЯ></reg>`
`  <src>~Mь <MR></src>`
`</w>`
	`<w reg=’мя’ src=’~Mь <MR>’>мь</w>`
	`<note type=’corr’>мʌ</note>`

`<c type=’punctuation’></c>`	`<pc></pc>`

Таблица 3.2. Предлагаемые замены тегов

положение вещей, вероятно, изменится: в настоящее время для СКАТ активно разрабатывается формат сюжетной разметки текстов корпуса [19].

Формальную разбивку документов на листы (`<div2>`) и страницы (`<div3>`) предлагается полностью заменить смысловой и для маркировки границ между ними пользоваться исключительно тегом `<pb/>` с обязательным атрибутом `@n`, обозначающим порядковый номер соответствующей страницы. Номера лицевых и оборотных сторон листа в соответствии с транслитерационными соглашениями СКАТ отличаются между собой по наличию при них специального префикса — дефиса.

Поскольку в режиме XTZ отсутствует поддержка специализированного тега-разделителя между колонками (`<cb/>`, column beginning), последние видится необходимым рассматривать как отдельные страницы и также отграни-

чивать друг от друга при помощи элемента <pb/>. При этом формат атрибута @n получает дополнительное расширение в виде суффикса a для первой колонки или b для второй. Отметим, что рукописи с тремя колонками и более чрезвычайно редки и в корпусе СКАТ не представлены, а рукопись с двумя колонками всего одна (житие Александра Свирского; РНБ, Пог. 874, XVI в.).

Замена элементов <l> на <lb/> осуществляется по аналогичному принципу; присваиваемые им порядковые номера являются простыми натуральными числами.

Направление преобразования тегов элементарных лексических единиц (<w>) было обосновано выше: в качестве их содержимого отныне выступает единственный вариант графического представления (совместимый с Unicode), а прочие конвертируются в одноимённые атрибуты. Для ошибочных написаний имеют место следующие спецификации: (1) в атрибут @reg записывается упрощённая форма исправленного варианта (ошибка не заносится); (2) внутрь тега <w> помещается Unicode-совместимое представление оригинального написания; (3) исправление обрамляется типизированным тегом <note>, непосредственно следующим за токеном. Далее это позволяет конструировать HTML-издания житийных текстов в первозданном виде, исправления же отображать как сноски.

Наконец, типизированный тег <c> заменён на <pc> в угоду краткости и соответствию нормам TEI [34, p. 575–577].

3.2.3. Обновление до Unicode 6.1

Выше было упомянуто, что стараниями В. А. Алексеева между собственной кодировкой исторических символов кириллицы, принятой в проекте СКАТ, и стандартом Unicode 5.1 было установлено практически полное взаимно однозначное соответствие. Исключение составляет ряд выносных

букв (ь, ы, у, u, и, i, w, а также е широкое и его йотированный аналог), к моменту окончания диссертационного исследования В. А. Алексеева не успевших войти в Unicode; однако им отмечалось, что предложение по внесению соответствующих дополнений в стандарт к тому времени уже было составлено и находилось на рассмотрении одной из рабочих групп ISO (International Organization for Standardization) [4, с. 21].

В обновлении Unicode до версии 6.1, увидевшем свет в январе 2012 г., данное предложение [25] было принято: все перечисленные выносные буквы стали доступны в составе блока Cyrillic Extended-B. Следовательно, отныне кодировать недостающие символы как сущности и приписывать им формальную декларацию нет необходимости: всем им были поставлены в соответствие их интернациональные эквиваленты. Кроме того, в таблицу символов шрифта AgioUnicode — совместимой со стандартом Unicode модификации кафедрального шрифта AGIO, разработанной В. А. Алексеевым [4, с. 49–50], — также были внесены соответствующие поправки.

3.2.4. Нормализация и лемматизация

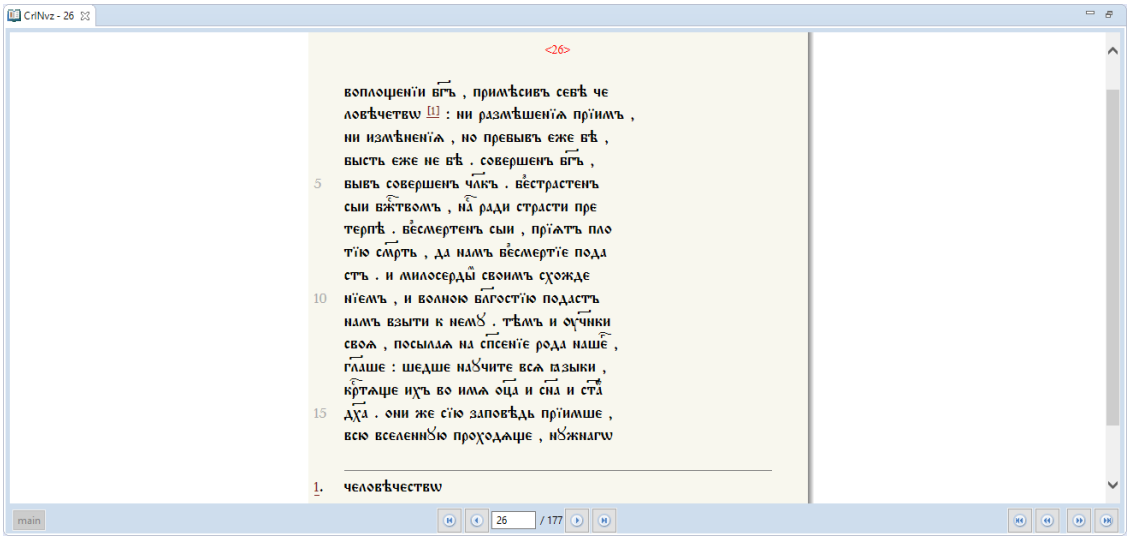
В обновлённое XML-представление были интегрированы все технологические наработки, составившие предмет обсуждения предыдущей главы. А именно: (1) в содержимое атрибута `@reg` словоформы отныне записываются не просто в упрощённой графике, но в нормализованном виде; (2) морфологически размеченные словоформы дополнительно снабжаются атрибутом `@ana`, где позиции разметки последовательно перечислены через точку с запятой; (3) леммы в случае их успешного определения попадают в атрибут `@lemma`.

Приложение Б иллюстрирует фрагмент XML-представления начального фрагмента жития Димитрия Прилуцкого.

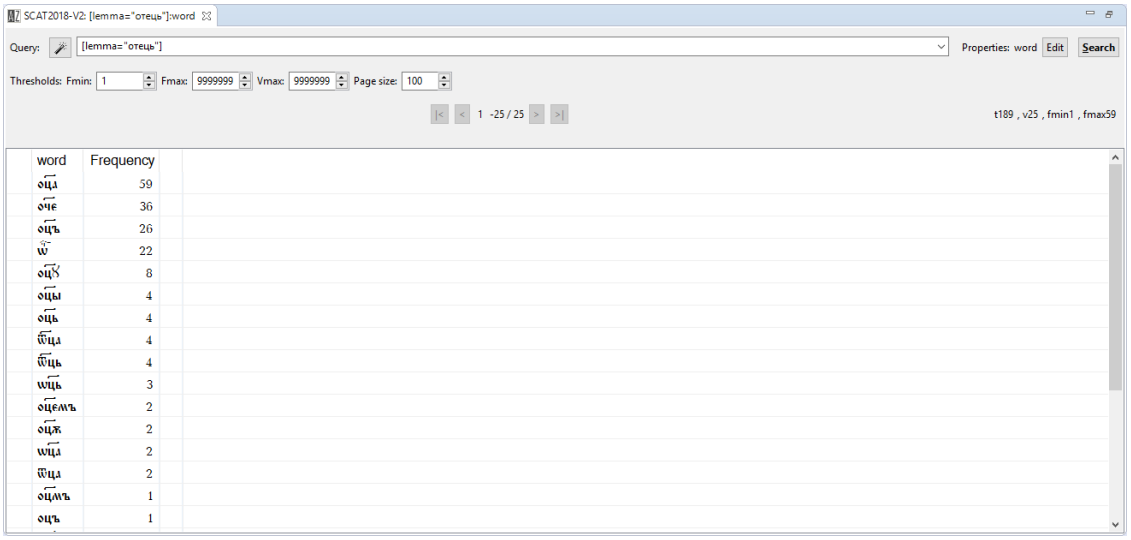
Выводы

На рис. 3.1 продемонстрированы основные функциональные возможности взаимодействия с корпусом СКАТ через платформу TXM: работа с загруженными рукописями в режиме чтения, построение запросов на языке CQL и получение по ним частотных списков и конкордансов, переход к полным контекстам искомых словоупотреблений.

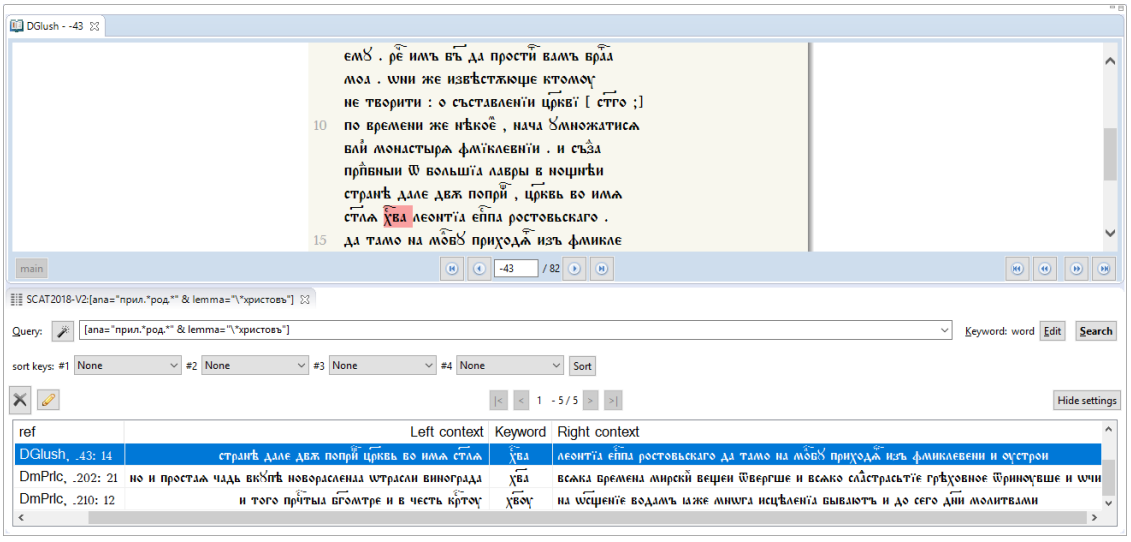
Возможность поиска по размеченным текстам на основании атрибутов `@ana` и `@lemma` наглядно свидетельствует о том, что отныне морфологическая разметка не автономна от основного корпуса, но непосредственно в него интегрирована, а задача лемматизации была нами решена не только локально-теоретически, но и глобально-практически: доступ к результатам работы реализованного алгоритма предоставляется любому, кто желает установить корпус на собственном оборудовании.



(*а*) Просмотр рукописи (КН 26)



(*б*) Частотный список по запросу на лемму отець



(*в*) Просмотр вхождения словоформы *х(с)ва (ДП 43 об./14)

Рис. 3.1. Работа с корпусом СКАТ на платформе ТХМ

Заключение

Итогом проделанной выпускной квалификационной работы стало решение следующих задач.

1. Был произведён обзор систем представления грамматических данных в существующих ныне восточнославянских исторических корпусах, в результате чего была обоснована актуальность проблемы лемматизации текстов в корпусе СКАТ для приведения последнего в соответствие с глобальным уровнем развития аналогичных проектов.

2. Были изучены основные трудности церковнославянского именного словоизменения, сопряжённые с задачей корректного определения леммы по заданной словоформе, и разработаны способы их формализации в ходе программной разработки алгоритма лемматизации морфологически размеченных житий.

3. Было усовершенствовано XML-представление текстов корпуса, что позволило далее загрузить их на платформу TXM, — не только выведя результаты работы алгоритма лемматизации на непосредственно практический уровень, но и расширив пользовательские возможности для практической работы с корпусом в целом.

Список литературы

1. *Аверина С. А.*, *Алексеева Е. Л.*, *Герд А. С.* Автоматизация обработки древних текстов // Прикладное языкознание : Учебник / под ред. А. С. Герда. — СПб., 1996. — С. 509–513.
2. *Аверина С. А.*, *Навтанович Л. М.*, *Попов М. Б.*, *Старовойтова О. А.* Старославянский язык : Учебник для высших учебных заведений Российской Федерации. — СПб., 2013.
3. *Азарова И. В.*, *Алексеева Е. Л.*, *Захарова Л. А.*, *Лемешев К. Н.*, *Биланчук Р. П.* Жития Феодосия Тотемского, Вассиана Тиксненского и Андрея Тотемского : Тексты и словоуказатель / под ред. А. С. Герда. — СПб., 2012.
4. *Алексеев В. А.* Расширение и реализация формата описания грамматических и графических данных корпуса СКАТ : Магистерская диссертация / Алексеев В. А. — СПбГУ, 2011.
5. *Алексеев В. А.*, *Алексеева Е. Л.*, *Касьяненко С. Е.* Грамматическая разметка в корпусе СКАТ // Труды международной конференции «Корпусная лингвистика — 2011». — СПб., 2011. — С. 69–73. — URL: https://events.spbu.ru/eventsContent/files/corpling/corpora2011/Alexeev_69.pdf (дата обр. 01.06.2018).
6. *Алексеева Е. Л.*, *Лаврентьев А. М.*, *Азарова И. В.*, *Захарова Л. А.* Разметка корпуса древнерусских агиографических текстов // Труды международной конференции «Корпусная лингвистика — 2004». — СПб., 2004. — С. 16–23. — URL: https://events.spbu.ru/eventsContent/files/corpling/corpora2004/Alexeeva_art.pdf (дата обр. 01.06.2018).

7. *Архангельский Т. А.*, *Мишина Е. И.*, *Пичхадзе А. А.* Система электронной грамматической разметки древнерусских и церковнославянских текстов // Palaeobulgarica / Старобългаристика. — София, 2014. — Т. 38, № 4. — С. 21–37.

8. *Баранов В. А.*, *Гулина О. В.*, *Миронов А. Н.* Морфологическая парадигма и её составляющие в системе «Манускрипт» // Информационные технологии и письменное наследие : Материалы III международной научной конференции (El'Manuscript-2010). — Уфа, Ижевск, 2010. — URL: https://textualheritage.org/ru/el-manusctipt-10/23.html (дата обр. 01.06.2018).

9. *Баранов В. А.*, *Миронов А. Н.*, *Лапин А. Н.* [и др.]. Автоматический морфологический анализатор древнерусского языка: лингвистические и технологические решения // 10-я юбилейная международная конференция «EVA 2007 Москва». — М., 2007. — URL: http://conf.evarussia.ru/upload/eva2007/reports/doklad_1318.pdf (дата обр. 01.06.2018).

10. *Гаврилова Т. С.*, *Шалганова Т. А.*, *Ляшевская О. Н.* К задаче автоматической лексико-грамматической разметки старорусского корпуса XV–XVII вв. // Вестник Православного Свято-Тихоновского гуманитарного университета. — 2016. — Т. 47, № 2. — С. 7–25. — URL: https://publications.hse.ru/articles/198589942 (дата обр. 01.06.2018).

11. *Герд А. С.*, *Азарова И. В.*, *Алексеева Е. Л.*, *Захарова Л. А.* Электронный корпус текстов по памятникам древнерусской агиографической литературы // Научно-техническая информация. — М., 2004. — Вып. 9. — С. 16–20. — URL: http://project.phil.spbu.ru/scat/document/nti.pdf (дата обр. 01.06.2018).

12. *Герд А. С.*, *Кузнецова Е. Л.*, *Аверина С. А.* [и др.]. Язык русской агиографии XVI века : Опыт автоматического анализа / под ред. А. С. Герда. — Л., 1990.

13. *Добрушина Е. Р.*, *Кравецкий А. Г.*, *Поляков А. Е.* Корпус и частотный грамматический корпусный словарь церковнославянского языка в составе Национального корпуса русского языка // Труды Института русского языка им. В. В. Виноградова. — М., 2015. — Вып. 6. — С. 116–141. — URL: https://elibrary.ru/item.asp?id=25390364 (дата обр. 01.06.2018).

14. *Иванова Е. С.* Схема разметки текста для электронной публикации древнерусских рукописей : Дипломное сочинение / Иванова Е. С. — СПбГУ, 2006.

15. *Иванова Т. А.* Старославянский язык. — СПб., 1998.

16. *Коваль С. А.* Лингвистические проблемы компьютерной морфологии. — СПб., 2005.

17. *Мишина Е. И.*, *Пичхадзе А. А.* Древнерусский подкорпус Национального корпуса русского языка // Труды Института русского языка им. В. В. Виноградова. — М., 2015. — Вып. 6. — С. 99–115. — URL: https://elibrary.ru/item.asp?id=25390363 (дата обр. 01.06.2018).

18. *Поляков А. Е.* Корпус церковнославянских текстов: проблемы орфографии и грамматики // Przegląd wschodnioeuropejski. — 2015. — Т. 5, № 1. — С. 245–254. — URL: http://www.ruslang.ru/doc/church-slav/conf4/05-polyakov.pdf (дата обр. 01.06.2018).

19. *Рогозина Е. А.* Уточнение и XML-разметка сюжетной схемы житий в корпусе агиографических текстов СКАТ // Структурная и прикладная

лингвистика. — СПб., 2015. — Вып. 11. — С. 168–173. — URL: http://elibrary.ru/item.asp?id=25849106 (дата обр. 01.06.2018).

20. *Сичинава Д. В.* Исторические корпуса Национального корпуса русского языка как инструмент диахронических исследований грамматики // Письменное наследие и информационные технологии : Материалы V международной научной конференции (El'Manuscript-2014). — София, Ижевск, 2014. — С. 226–229. — URL: https://textualheritage.org/ru/el-manusctipt-2014/48.html (дата обр. 01.06.2018).

21. *Срезневский И. И.* Материалы для словаря древнерусского языка по письменным памятникам : в 3 т. — Санкт-Петербург, 1893–1912.

22. *Уфлянд Е. Г.* Автоматическое сведение орфографических вариантов словоформ в электронном корпусе текстов по памятникам агиографической литературы 16–17 веков «СКАТ» : Дипломное сочинение / Уфлянд Е. Г. — СПбГУ, 2008.

23. *Уфлянд Е. Г.*, *Алексеева Е. Л.* Сокращение вариативности написания словоформ в служебных компонентах агиографического корпуса СКАТ // Труды международной конференции «Корпусная лингвистика — 2008». — СПб., 2008. — С. 376–378. — URL: https://events.spbu.ru/eventsContent/files/corpling/corpora2008/UflandAlexeeva_376_378.pdf (дата обр. 01.06.2018).

24. *Хабургаев Г. А.* Старославянский язык. — М., 1986.

25. *Everson M.*, *Baranov V.*, *Miklas H.*, *Rabus A.* Proposal to Encode Nine Cyrillic Characters for Slavonic / UC Berkeley Script Encoding Initiative (Universal Scripts Project). — 2010. — URL: http://www.unicode.org/L2/L2010/10002-n3748-cyrillic-superscripts.pdf (visited on 06/01/2018).

26. *Heiden S.* The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme // 24th Pacific Asia Conference on Language, Information and Computation. — Sendai, 2010. — P. 389–398. — URL: https://halshs.archives-ouvertes.fr/halshs-00549764 (visited on 06/01/2018).

27. *Krauwer S.* The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap // Proceedings of the International Workshop *Speech and Computer* (SPECOM). — Moscow, 2003. — P. 8–15. — URL: http://www.elsnet.org/dox/krauwer-specom2003.pdf (visited on 06/01/2018).

28. Manuel de TXM / ENS, Lyon & Université de Franche-Comté. — Version 0.7.9. — 16/05/2018. — URL : http://textometrie.ens-lyon.fr/files/documentation/Manuel%20de%20TXM%200.7%20FR.pdf (visité le 01/06/2018).

29. *Meyer R.* New Wine in Old Wineskins?—Tagging Old Russian via Annotation Projection from Modern Translations // Russian Linguistics. — 2011. — Vol. 35, issue 2. — P. 267–281.

30. *Meyer R.* The History of Null Subjects in North Slavonic. A Corpus-Based Diachronic Investigation : Habilitation Thesis / Meyer Roland. — University of Regensburg, 2012.

31. *Mitrenina O.* The Corpora of Old and Middle Russian Texts as an Advanced Tool for Exploring an Extinguished Language // Scrinium : Journal of Patrology, Critical Hagiography, and Ecclesiastical History. — 2014. — Vol. 10, no. 1. — P. 455–461. — URL: https://www.academia.edu/6923912 (visited on 06/01/2018).

32. *Passarotti M.* Leaving Behind the Less-Resourced Status. The Case of Latin through the Experience of the *Index Thomisticus* Treebank // 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages (LREC). — Valetta, 2010. — P. 27–32. — URL: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W21.pdf (visited on 06/01/2018).

33. *Porter M. F.* An Algorithm for Suffix Stripping // Program : Electronic Library and Information Systems. — 1980. — Vol. 14, no. 3. — P. 130–137. — URL: https://tartarus.org/martin/PorterStemmer/def.txt (visited on 06/01/2018).

34. TEI P5 : Guidelines for Electronic Text Encoding and Interchange / The TEI Consortium. — Version 3.3.0. — 01/31/2018. — URL: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf (visited on 06/01/2018).

35. *Waldenfels R. von*, *Rabus A.* Recycling the Metropolitan : Building an Electronic Corpus on the Basis of the Edition of the *Velikie Minei Čet'i* // Scripta & e-Scripta. — 2015. — Issue 14/15. — P. 27–38. — URL: https://www.academia.edu/18777247 (visited on 06/01/2018).

Приложение А

Пример морфологической разметки жития Димитрия Прилуцкого

м(с)ца	сущ	jo	род	ед	м	
февраlr.	сущ	jo	род	ед	м	
аi#.	11					
житiе	сущ	jo	им	ед	ср	
прпw(до)&бнаго	прил	тв	род	ед	м	
w(т)ца	сущ	jo	род	ед	м	
ншего#	мест	м	род	ед	м	
*дiмитрiа,	сущ	jo	род	ед	м	
игu&мена	сущ	o	род	ед	м	
*прилuцкаго.	прил	тв	род	ед	м	
*вологоцкагw&	прил	тв	род	ед	м	
чюдотворца.	сущ	jo	род	ед	м	
творенiе	сущ	jo	им	ед	ср	
тоа	мест	тв	род	ед	ж	
же	част					
w&бители,	сущ	i	род	ед	ж	
игumена	сущ	o	род	ед	м	
*макарiа.	сущ	jo	род	ед	м	
бл(с)ви	гл	повел	2	ед	4	
w(ч);&	сущ	jo/o	зв	ед	м	
стое#	прил	тв	вин	ед	ср	
житiе	сущ	jo	вин	ед	ср	
пожившu(х),	прич	м	прош	род	мн	м
сты(х)#	прил	тв	род	мн	м	
препо&добны(х)	прил	тв	род	мн	м	
muжеи	сущ	jo/i	род	мн	м	

ВЕЛИКЫ(Х).	прил	тв	род	мн	м	
ИЖЕ	мест	м	им	мн	м	
В	пред					
ПО&СТНЫ(Х)	прил	тв	мест	мн	м	
ПОДВИЗЪ(Х)	сущ	о	мест	мн	м	*
ПРОСІАВШИ(Х),	прич	м	прош	род	мн	м
И	союз					
В	пред					
НЕ&ПРЕСТАННЫ(Х)	прил	тв	мест	мн	ж	
МЛТВА(Х)#	сущ	а	мест	мн	ж	
КЪ	пред					
БU#.	сущ	о	дат	ед	м	
ВЕЛИКD&	прил	а	вин	ед	ж	
ПОБѢДU	сущ	а	вин	ед	ж	
НА	пред					
ВРАГЫ	сущ	о	вин	мн	м	
ПОКАЗАВШЕ.	прич	jo/en	прош	им	мн	м
КРА&СНАА	прил	тв	вин	мн	ср	
ЖЕ	част					
И	союз					
CUETHAA	прил	тв	вин	мн	ср	
МИРА	сущ	о	род	ед	м	
СЕГО	мест	м	род	ед	м	
W(Т)ВЕРГЬ&ШИ(Х).	прич	м	прош	род	мн	м
БUДUЩАА	прич	м	наст	род	ед	ж
РАДИ	посл					
ЖИЗНИ.	сущ	і	род	ед	ж	
ЯЖЕ&	мест	м	вин	мн	ср	
БЪ#	сущ	о	им	ед	м	
UГОТОВА	гл	изъяв	аор гл	3	ед	
ЛЮБRЩИМЪ	прич	м	наст	дат	мн	м

ЕГО.	мест	м	вин/род	ед	м	
АЩЕ&	союз					
БО	союз					
ЕЛЛИНСТIИ	прил	тв	им	мн	м	*
БАСНОТВОРЦИ.	сущ	jo	им	мн	м	
И	союз					
ИЖЕ&	мест	м	им	мн	м	
С	пред					
НИМИ	мест	м	тв	мн	м	
НЕЧЕСТИВIИ	прил	тв	им	мн	м	
RЗЫЦИ.	сущ	o	им	мн	м	*
НЕ	част					
ЗНА&ЮЩИ	прич	jo	наст	им	мн	м
БА#	сущ	o	вин/род	ед	м	
ТВОРЦА	сущ	jo	вин/род	ед	м	
ВСRЧЕСКИМЪ.	прил	тв	дат	мн	ср	
ПРА&З(Д)НИКИ	сущ	o	вин	мн	м	
И	союз					
ПОКЛОНЕНIЕ	сущ	jo	вин	ед	ср	
ИДОЛОМЪ	сущ	o	дат	мн	м	
ПРИНО&СRЩЕ.	прич	jo/en	наст	им	мн	м
ЕЛИКО	мест	тв	вин	ед	ср	
КОЖ(Д)О	мест	тв	им	ед	м	
ИХЪ	мест	м	род	мн	м	
МОЖАAXU&	гл	изъяв	имп	3	мн	
ТЩАЩЕСR,	прич/в	en	наст	им	мн	м
W(Т)	пред					
ИМѢНIИ	сущ	jo	род	мн	ср	
СВОИ(Х).	мест	м	род	мн	ср	
ТѢ(М)&	мест	тв	дат	мн	м	
БЕЗ	пред					

UMA	сущ	о	род	ед	м	
ПРЕ(Д)ЛАГАХU.	гл	изъяв	имп	3	мн	
КОЛМИ	нар					
ЖЕ	част					
НА(М).&	мест	личн	1	дат	мн	
ИЖЕ	мест	м	им	мн	м	
БЖ(С)ТВЕНОЮ	прил	тв	тв	ед	ж	
БЛГОДАТIЮ#	сущ	i	тв	ед	ж	
W(Т)	пред					
ГА#	сущ	i/o	род	ед	м	
БА#&	сущ	о	род	ед	м	
ПОСЂЩЕНЫМЪ.	прич	тв	прош	дат	мн	м
ПОДОБАЕТЬ	гл	изъяв	н/б	3	ед	3
ДОСТО&ИНО	нар					
ПАМRТИ	сущ	i	вин	мн	ж	
СТЫ(Х)#	прил	тв	род	мн	м	
W(Т)ЦЬ#	сущ	jo	род	мн	м	
ДХОВНО#	нар					
ПРАZ -201 ЗНОВАТИ.	инф					
И	союз					
W(Т)	пред					
*ХА#	сущ	о	род	ед	м	
БА#	сущ	о	род	ед	м	
ДАРОВАННАА	прич	тв	прош	вин	мн	ср
ТЂ(М),&	мест	тв	дат	мн	м	
ЧЮДЕСА	сущ	es	вин	мн	ср	
ПОХВАЛИТИ	инф					
ВЪ	пред					
QАЛМЂ(Х)	сущ	о	мест	мн	м	+о
И	союз					
ПЂ&НIИХЪ	сущ	jo	мест	мн	ср	

СЛGЖАЩЕ	прич	jo/en	наст	им	мн	м
ГВИ#.	сущ	i/u	дат	ед	м	
ПАMRТИ	сущ	i	род	ед	ж	
И(Х)&	мест	м	род	мн	м	
РАДИ	посл					
ПОДОБАЕТЪ	гл	изъяв	н/б	3	ед	3
ЯЖЕ	мест	м	вин	мн	ср	
ВИДѢХОМЪ.&	гл	изъяв	аор гл	1	мн	
И	союз					
АЩЕ	союз					
И	союз					
ПРЕЖНRА	прил	м	вин	мн	ср	
И	союз					
СЛЫШАХОМЪ.	гл	изъяв	аор гл	1	мн	
ПИ&САНІЮ	сущ	jo	дат	ед	ср	
ИСТИННАА	прил	тв	вин	мн	ср	
ПОЛОЖИТИ.	инф					

Приложение Б

Пример обновлённого XML-представления жития Димитрия Прилуцкого

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Житие Димитрия Прилуцкого</title>
        <respStmt>
          <resp>редактор</resp>
          <name>А. С. Герд</name>
        </respStmt>
        <respStmt>
          <resp>составитель</resp>
          <name>И. В. Азарова</name>
        </respStmt>
        <respStmt>
          <resp>составитель</resp>
          <name>Е. Л. Алексеева</name>
        </respStmt>
        <respStmt>
          <resp>составитель</resp>
          <name>Л. А. Захарова</name>
        </respStmt>
        <respStmt>
          <resp>составитель</resp>
          <name>К. Н. Лемешев</name>
        </respStmt>
        <respStmt>
          <resp>конвертация в формат XML-TEI</resp>
          <name>К. В. Сипунин</name>
        </respStmt>
      </titleStmt>
      <publicationStmt>
        <publisher>Издательство Санкт-Петербургского университета</publisher>
        <pubPlace>Санкт-Петербург</pubPlace>
        <date>2003</date>
        <idno type="ISBN">5-288-03308-0</idno>
      </publicationStmt>
      <sourceDesc>
        <bibl>РНБ, Соф. 1361; XVI в.</bibl>
      </sourceDesc>
    </fileDesc>
```

```
</teiHeader>
<text><body><ab>
  <pb n="201"/><lb n="1"/>
  <w xml:id="DmPrlc.1" ana="сущ;jo;род;ед;м" lemma="месяць" reg="месяца" src="М(с)ЦА">м□□ца</w>
  <w xml:id="DmPrlc.2" ana="сущ;jo;род;ед;м" lemma="февраль" reg="февраля" src="FEВРАЛR">ѳевралѧ</w>
  <pc xml:id="DmPrlc.3">.</pc>
  <num><w xml:id="DmPrlc.4" reg="11" src="AI#">аїꙋ</w></num>
  <pc xml:id="DmPrlc.5">.</pc>
  <w xml:id="DmPrlc.6" ana="сущ;jo;им;ед;ср" lemma="житие" reg="житие" src="ЖИТIЕ">житїе</w>
  <w xml:id="DmPrlc.7" ana="прил;тв;род;ед;м" lemma="преподобныи" reg="преподобнаго"
  ↪ src="ПРПW(ДО)&amp;БНАГО">прпw□□<lb n="2"/>бнаго</w>
  <w xml:id="DmPrlc.8" ana="сущ;jo;род;ед;м" lemma="отець" reg="о(т)ца" src="W(Т)ЦА">w̃ца</w>
  <w xml:id="DmPrlc.9" ana="мест;м;род;ед;м" lemma="нашь" reg="нашего" src="НШЕГО#">нш̃его</w>
  <name><w xml:id="DmPrlc.10" ana="сущ;jo;род;ед;м" lemma="*димитрии" reg="*димитрия"
  ↪ src="*ДIМИТРIА">дїмитрїа</w></name>
  <pc xml:id="DmPrlc.11">,</pc>
  <w xml:id="DmPrlc.12" ana="сущ;o;род;ед;м" lemma="игуменъ" reg="игумена" src="ИГU&amp;МЕНА">игѹ<lb
  ↪ n="3"/>мена</w>
  <name><w xml:id="DmPrlc.13" ana="прил;тв;род;ед;м" lemma="*прилуцкии" reg="*прилуцкаго"
  ↪ src="*ПРИЛUЦКАГО">прилѹцкаго</w></name>
  <pc xml:id="DmPrlc.14">.</pc>
  <name><w xml:id="DmPrlc.15" ana="прил;тв;род;ед;м" lemma="*вологоцкии" reg="*вологоцкаго"
  ↪ src="*ВОЛОГОЦКАГW">вологоцкагw</w></name>
  <lb n="4"/>
  <w xml:id="DmPrlc.16" ana="сущ;jo;род;ед;м" lemma="чюдотворець" reg="чюдотворца"
  ↪ src="ЧЮДОТВОРЦА">чюдотворца</w>
  <pc xml:id="DmPrlc.17">.</pc>
  <w xml:id="DmPrlc.18" ana="сущ;jo;им;ед;ср" lemma="творение" reg="творение"
  ↪ src="ТВОРЕНIЕ">творенїе</w>
  <w xml:id="DmPrlc.19" ana="мест;тв;род;ед;ж" lemma="тои" reg="тоя" src="ТОА">тоа</w>
  <w xml:id="DmPrlc.20" ana="част" lemma="же" reg="же" src="ЖЕ">же</w>
  <w xml:id="DmPrlc.21" ana="сущ;i;род;ед;ж" lemma="обитель" reg="обители" src="W&amp;БИТЕЛИ">w<lb
  ↪ n="5"/>бители</w>
  <pc xml:id="DmPrlc.22">,</pc>
  <w xml:id="DmPrlc.23" ana="сущ;o;род;ед;м" lemma="игуменъ" reg="игумена" src="ИГUМЕНА">игѹмена</w>
  <name><w xml:id="DmPrlc.24" ana="сущ;jo;род;ед;м" lemma="*макарии" reg="*макария"
  ↪ src="*МАКАРIА">макарїа</w></name>
  <pc xml:id="DmPrlc.25">.</pc>
  <w xml:id="DmPrlc.26" ana="гл;повел;2;ед;4" reg="благослови" src="БЛ(с)ВИ">бл□□ви</w>
  <w xml:id="DmPrlc.27" ana="сущ;jo/o;зв;ед;м" lemma="отець" reg="о(ч)" src="W(ч)">w□□</w>
  <pc xml:id="DmPrlc.28">;</pc>
  <lb n="6"/>
  <w xml:id="DmPrlc.29" ana="прил;тв;вин;ед;ср" lemma="святыи" reg="святое" src="СТОЕ#">стѣе</w>
  <w xml:id="DmPrlc.30" ana="сущ;jo;вин;ед;ср" lemma="житие" reg="житие" src="ЖИТIЕ">житїе</w>
  <w xml:id="DmPrlc.31" ana="прич;м;прош;род;мн;м" lemma="пожити" reg="пожившихъ"
  ↪ src="ПОЖИВШИ(Х)">пожившꙋ</w>
  <pc xml:id="DmPrlc.32">,</pc>
  <w xml:id="DmPrlc.33" ana="прил;тв;род;мн;м" lemma="святыи" reg="святыхъ" src="СТЫ(Х)#">сты̑□</w>
```

```
<w xml:id="DmPrlc.34" ana="прил;тв;род;мн;м" lemma="преподобныи" reg="преподобныхъ"
↪ src="ПРЕПО&amp;ДОБНЫ(Х)">препо<lb n="7"/>добны</w>
<w xml:id="DmPrlc.35" ana="сущ;jo/i;род;мн;м" lemma="мужь" reg="мужеи" src="MUЖЕИ">мꙋжеи</w>
<w xml:id="DmPrlc.36" ana="прил;тв;род;мн;м" lemma="великии" reg="великихъ"
↪ src="ВЕЛИКЫ(Х)">великы</w>
<pc xml:id="DmPrlc.37">.</pc>
<w xml:id="DmPrlc.38" ana="мест;м;им;мн;м" lemma="иже" reg="иже" src="ИЖЕ">иже</w>
<w xml:id="DmPrlc.39" ana="пред" lemma="въ" reg="въ" src="В">в</w>
<w xml:id="DmPrlc.40" ana="прил;тв;мест;мн;м" lemma="постныи" reg="постныхъ"
↪ src="ПО&amp;СТНЫ(Х)">по<lb n="8"/>стны</w>
<w xml:id="DmPrlc.41" ana="сущ;о;мест;мн;м;*" lemma="подвигъ" reg="подвиз+хъ"
↪ src="ПОДВИЗ+(Х)">подвизѣ</w>
<w xml:id="DmPrlc.42" ana="прич;м;прош;род;мн;м" lemma="просияти" reg="просиявшихъ"
↪ src="ПРОСIАВШИ(Х)">просїавши</w>
<pc xml:id="DmPrlc.43">,</pc>
<w xml:id="DmPrlc.44" ana="союз" lemma="и" reg="и" src="И">и</w>
<w xml:id="DmPrlc.45" ana="пред" lemma="въ" reg="въ" src="В">в</w>
<w xml:id="DmPrlc.46" ana="прил;тв;мест;мн;ж" lemma="непрестанныи" reg="непрестанныхъ"
↪ src="НЕ&amp;ПРЕСТАННЫ(Х)">не<lb n="9"/>престанны</w>
<w xml:id="DmPrlc.47" ana="сущ;а;мест;мн;ж" lemma="молитва" reg="молитвахъ" src="МЛТВА(Х)#">млт҃ва</w>
<w xml:id="DmPrlc.48" ana="пред" lemma="къ" reg="къ" src="КЪ">къ</w>
<w xml:id="DmPrlc.49" ana="сущ;о;дат;ед;м" lemma="богъ" reg="богу" src="БU#">бꙋ҃</w>
<pc xml:id="DmPrlc.50">.</pc>
<w xml:id="DmPrlc.51" ana="прил;а;вин;ед;ж" lemma="великии" reg="велику" src="ВЕЛИKD">велик</w>
<lb n="10"/>
<w xml:id="DmPrlc.52" ana="сущ;а;вин;ед;ж" lemma="поб+да" reg="поб+ду" src="ПОБ+ДU">побѣдꙋ</w>
<w xml:id="DmPrlc.53" ana="пред" lemma="на" reg="на" src="НА">на</w>
<w xml:id="DmPrlc.54" ana="сущ;о;вин;мн;м" lemma="врагъ" reg="враги" src="ВРАГЫ">врагы</w>
<w xml:id="DmPrlc.55" ana="прич;jo/en;прош;им;мн;м" lemma="показати" reg="показавше"
↪ src="ПОКАЗАВШЕ">показавше</w>
<pc xml:id="DmPrlc.56">.</pc>
<w xml:id="DmPrlc.57" ana="прил;тв;вин;мн;ср" lemma="красныи" reg="красная" src="КРА&amp;СНАА">кра<lb
↪ n="11"/>снаа</w>
<w xml:id="DmPrlc.58" ana="част" lemma="же" reg="же" src="ЖЕ">же</w>
<w xml:id="DmPrlc.59" ana="союз" lemma="и" reg="и" src="И">и</w>
<w xml:id="DmPrlc.60" ana="прил;тв;вин;мн;ср" lemma="суетныи" reg="суетная" src="CUETHAA">сꙋетнаа</w>
<w xml:id="DmPrlc.61" ana="сущ;о;род;ед;м" lemma="миръ" reg="мира" src="МИРА">мира</w>
<w xml:id="DmPrlc.62" ana="мест;м;род;ед;м" lemma="сеи" reg="сего" src="СЕГО">сего</w>
<w xml:id="DmPrlc.63" ana="прич;м;прош;род;мн;м" lemma="отвергнути" reg="о(т)вергшихъ"
↪ src="W(Т)ВЕРГЬ&amp;ШИ(Х)">w̃вергь<lb n="12"/>ши</w>
<pc xml:id="DmPrlc.64">.</pc>
<w xml:id="DmPrlc.65" ana="прич;м;наст;род;ед;ж" reg="будущая" src="БUДUЩАА">бꙋдꙋщаа</w>
<w xml:id="DmPrlc.66" ana="посл" lemma="ради" reg="ради" src="РАДИ">ради</w>
<w xml:id="DmPrlc.67" ana="сущ;i;род;ед;ж" lemma="жизнь" reg="жизни" src="ЖИЗНИ">жизни</w>
<pc xml:id="DmPrlc.68">.</pc>
<w xml:id="DmPrlc.69" ana="мест;м;вин;мн;ср" lemma="иже" reg="яже" src="ЯЖЕ">же</w>
<lb n="13"/>
<w xml:id="DmPrlc.70" ana="сущ;о;им;ед;м" lemma="богъ" reg="богъ" src="БЪ#">бъ҃</w>
```

```
<w xml:id="DmPrlc.71" ana="гл;изъяв;аор гл;3;ед" lemma="уготовати" reg="уготова"
↪ src="UГОТОВА">ѹготова</w>
<w xml:id="DmPrlc.72" ana="прич;м;наст;дат;мн;м" reg="любящимъ" src="ЛЮБRЩИМЪ">любѧщимъ</w>
<w xml:id="DmPrlc.73" ana="мест;м;вин/род;ед;м" lemma="и" reg="его" src="ЕГО">его</w>
<pc xml:id="DmPrlc.74">.</pc>
<w xml:id="DmPrlc.75" ana="союз" lemma="аще" reg="аще" src="АЩЕ">аще</w>
<lb n="14"/>
<w xml:id="DmPrlc.76" ana="союз" lemma="бо" reg="бо" src="БО">бо</w>
<w xml:id="DmPrlc.77" ana="прил;тв;им;мн;м;*" lemma="еллинскии" reg="еллинстии"
↪ src="ЕЛЛИНСТIИ">еллинстїи</w>
<w xml:id="DmPrlc.78" ana="сущ;jo;им;мн;м" lemma="баснотворець" reg="баснотворцы"
↪ src="БАСНОТВОРЦИ">баснотворци</w>
<pc xml:id="DmPrlc.79">.</pc>
<w xml:id="DmPrlc.80" ana="союз" lemma="и" reg="и" src="И">и</w>
<w xml:id="DmPrlc.81" ana="мест;м;им;мн;м" lemma="иже" reg="иже" src="ИЖЕ">иже</w>
<lb n="15"/>
<w xml:id="DmPrlc.82" ana="пред" lemma="съ" reg="съ" src="С">с</w>
<w xml:id="DmPrlc.83" ana="мест;м;тв;мн;м" lemma="и" reg="ними" src="НИМИ">ними</w>
<w xml:id="DmPrlc.84" ana="прил;тв;им;мн;м" lemma="нечестивыи" reg="нечестивии"
↪ src="НЕЧЕСТИВIИ">нечестивїи</w>
<w xml:id="DmPrlc.85" ana="сущ;o;им;мн;м;*" lemma="языкъ" reg="языцы" src="RЗЫЦИ">ѧзыци</w>
<pc xml:id="DmPrlc.86">.</pc>
<w xml:id="DmPrlc.87" ana="част" lemma="не" reg="не" src="НЕ">не</w>
<w xml:id="DmPrlc.88" ana="прич;jo;наст;им;мн;м" reg="знающи" src="ЗНА&amp;ЮЩИ">зна<lb n="16"/>ющи</w>
<w xml:id="DmPrlc.89" ana="сущ;o;вин/род;ед;м" lemma="богъ" reg="бога" src="БА#">ба҃</w>
<w xml:id="DmPrlc.90" ana="сущ;jo;вин/род;ед;м" lemma="творець" reg="творца" src="ТВОРЦА">творца</w>
<w xml:id="DmPrlc.91" ana="прил;тв;дат;мн;ср" lemma="всяческии" reg="всяческимъ"
↪ src="ВСRЧЕСКИМЪ">всѧческимъ</w>
<pc xml:id="DmPrlc.92">.</pc>
<w xml:id="DmPrlc.93" ana="сущ;o;вин;мн;м" lemma="праздникъ" reg="праз(д)ники"
↪ src="ПРА&amp;З(Д)НИКИ">пра<lb n="17"/>з⃞ники</w>
<w xml:id="DmPrlc.94" ana="союз" lemma="и" reg="и" src="И">и</w>
<w xml:id="DmPrlc.95" ana="сущ;jo;вин;ед;ср" lemma="поклонение" reg="поклонение"
↪ src="ПОКЛОНЕНIЕ">поклоненїе</w>
<w xml:id="DmPrlc.96" ana="сущ;o;дат;мн;м" lemma="идолъ" reg="идоломъ" src="ИДОЛОМЪ">идоломъ</w>
<w xml:id="DmPrlc.97" ana="прич;jo/en;наст;им;мн;м" reg="приносяще" src="ПРИНО&amp;CRЩЕ">прино<lb
↪ n="18"/>сѧще</w>
<pc xml:id="DmPrlc.98">.</pc>
<w xml:id="DmPrlc.99" ana="мест;тв;вин;ед;ср" lemma="елико" reg="елико" src="ЕЛИКО">елико</w>
<w xml:id="DmPrlc.100" ana="мест;тв;им;ед;м" lemma="кождо" reg="кож(д)о" src="КОЖ(Д)О">кож⃞о</w>
<w xml:id="DmPrlc.101" ana="мест;м;род;мн;м" lemma="и" reg="ихъ" src="ИХЪ">ихъ</w>
<w xml:id="DmPrlc.102" ana="гл;изъяв;имп;3;мн" reg="можаху" src="МОЖААХU">можаахѹ</w>
<lb n="19"/>
<w xml:id="DmPrlc.103" ana="прич/в;en;наст;им;мн;м" reg="тщащеся" src="ТЩАЩЕСR">тщащесѧ</w>
<pc xml:id="DmPrlc.104">,</pc>
<w xml:id="DmPrlc.105" ana="пред" lemma="отъ" reg="о(т)" src="W(Т)">ѿ</w>
<w xml:id="DmPrlc.106" ana="сущ;jo;род;мн;ср" lemma="им+ние" reg="им+нии" src="ИМ+НIИ">имѣнїи</w>
<w xml:id="DmPrlc.107" ana="мест;м;род;мн;ср" lemma="свои" reg="своихъ" src="СВОИ(Х)">свои⃞</w>
<pc xml:id="DmPrlc.108">.</pc>
```

```
<w xml:id="DmPrlc.109" ana="мест;тв;дат;мн;м" lemma="тои" reg="т+мь" src="Т+(М)">тѣ</w>
<lb n="20"/>
<w xml:id="DmPrlc.110" ana="пред" lemma="безъ" reg="безъ" src="БЕЗ">без</w>
<w xml:id="DmPrlc.111" ana="сущ;о;род;ед;м" lemma="умь" reg="ума" src="UMA">ѹма</w>
<w xml:id="DmPrlc.112" ana="гл;изъяв;имп;3;мн" reg="пре(д)лагаху" src="ПРЕ(Д)ЛАГАХU">пре лагахѹ</w>
<pc xml:id="DmPrlc.113">.</pc>
<w xml:id="DmPrlc.114" ana="нар" lemma="колми" reg="колми" src="КОЛМИ">колми</w>
<w xml:id="DmPrlc.115" ana="част" lemma="же" reg="же" src="ЖЕ">же</w>
<w xml:id="DmPrlc.116" ana="мест;личн;1;дат;мн" lemma="мы" reg="намъ" src="НА(М)">на</w>
<pc xml:id="DmPrlc.117">.</pc>
<lb n="21"/>
<w xml:id="DmPrlc.118" ana="мест;м;им;мн;м" lemma="иже" reg="иже" src="ИЖЕ">иже</w>
<w xml:id="DmPrlc.119" ana="прил;тв;тв;ед;ж" lemma="божественыи" reg="божественою"
↪ src="БЖ(с)ТВЕНОЮ">бж твеною</w>
<w xml:id="DmPrlc.120" ana="сущ;i;тв;ед;ж" lemma="благодать" reg="благодатию"
↪ src="БЛГОДАТІЮ#">блґодатїю</w>
<w xml:id="DmPrlc.121" ana="пред" lemma="отъ" reg="о(т)" src="W(Т)">ѿ</w>
<w xml:id="DmPrlc.122" ana="сущ;i/o;род;ед;м" lemma="господь" reg="господа" src="ГА#">гаꙋ</w>
<w xml:id="DmPrlc.123" ana="сущ;о;род;ед;м" lemma="богъ" reg="бога" src="БА#">баꙋ</w>
<lb n="22"/>
<w xml:id="DmPrlc.124" ana="прич;тв;прош;дат;мн;м" lemma="пос+тити" reg="пос+щенымъ"
↪ src="ПОС+ЩЕНЫМЪ">посѣщенымъ</w>
<pc xml:id="DmPrlc.125">.</pc>
<w xml:id="DmPrlc.126" ana="гл;изъяв;н/б;3;ед;3" reg="подобаеть" src="ПОДОБАЕТЬ">подобаеть</w>
<w xml:id="DmPrlc.127" ana="нар" lemma="достоино" reg="достоино" src="ДОСТО&amp;ИНО">досто<lb
↪ n="23"/>ино</w>
<w xml:id="DmPrlc.128" ana="сущ;i;вин;мн;ж" lemma="память" reg="памяти" src="ПАМЯТИ">памѧти</w>
<w xml:id="DmPrlc.129" ana="прил;тв;род;мн;м" lemma="святыи" reg="святыхъ" src="СТЫ(Х)#">стꙑ</w>
<w xml:id="DmPrlc.130" ana="сущ;jo;род;мн;м" lemma="отець" reg="отець" src="W(Т)ЦЬ#">ѿць</w>
<w xml:id="DmPrlc.131" ana="нар" lemma="духовно" reg="духовно" src="ДХОВНО#">дхо҃вно</w>
<w xml:id="DmPrlc.132" ana="инф" lemma="празновати" reg="празновати" src="ПРАZ -201 ЗНОВАТИ">пра<pb
↪ n="-201"/><lb n="1"/>зновати</w>
<w xml:id="DmPrlc.133" ana="союз" lemma="и" reg="и" src="И">и</w>
<w xml:id="DmPrlc.134" ana="пред" lemma="отъ" reg="о(т)" src="W(Т)">ѿ</w>
<name><w xml:id="DmPrlc.135" ana="сущ;о;род;ед;м" lemma="*христосъ" reg="*христа"
↪ src="*ХА#">хаꙋ</w></name>
<w xml:id="DmPrlc.136" ana="сущ;о;род;ед;м" lemma="богъ" reg="бога" src="БА#">баꙋ</w>
<w xml:id="DmPrlc.137" ana="прич;тв;прош;вин;мн;ср" lemma="даровати" reg="дарованная"
↪ src="ДАРОВАННАА">дарованнаа</w>
<w xml:id="DmPrlc.138" ana="мест;тв;дат;мн;м" lemma="тои" reg="т+мь" src="Т+(М)">тѣ</w>
<pc xml:id="DmPrlc.139">,</pc>
<lb n="2"/>
<w xml:id="DmPrlc.140" ana="сущ;es;вин;мн;ср" lemma="чюдо" reg="чюдеса" src="ЧЮДЕСА">чюдеса</w>
<w xml:id="DmPrlc.141" ana="инф" lemma="похвалити" reg="похвалити" src="ПОХВАЛИТИ">похвалити</w>
<w xml:id="DmPrlc.142" ana="пред" lemma="въ" reg="въ" src="ВЪ">въ</w>
<w xml:id="DmPrlc.143" ana="сущ;о;мест;мн;м;+о" lemma="псаломъ" reg="псалм+хъ"
↪ src="QАЛМ+(Х)">ѱалмѣ</w>
<w xml:id="DmPrlc.144" ana="союз" lemma="и" reg="и" src="И">и</w>
```

```
200       <w xml:id="DmPrlc.145" ana="сущ;jo;мест;мн;ср" lemma="п+ние" reg="п+ниихъ" src="П+&amp;HIИХЪ">пѣ<lb
          ↪  n="3"/>нїихъ</w>
201       <w xml:id="DmPrlc.146" ana="прич;jo/en;наст;им;мн;м" reg="служаще" src="СЛGЖАЩЕ">слѫжаще</w>
202       <w xml:id="DmPrlc.147" ana="сущ;i/u;дат;ед;м" lemma="господь" reg="господеви" src="ГВИ#">гвй</w>
203       <pc xml:id="DmPrlc.148">.</pc>
204       <w xml:id="DmPrlc.149" ana="сущ;i;род;ед;ж" lemma="память" reg="памяти" src="ПАМRТИ">памѧти</w>
205       <w xml:id="DmPrlc.150" ana="мест;м;род;мн;м" lemma="и" reg="ихъ" src="И(Х)">и</w>
206       <lb n="4"/>
207       <w xml:id="DmPrlc.151" ana="посл" lemma="ради" reg="ради" src="РАДИ">ради</w>
208       <w xml:id="DmPrlc.152" ana="гл;изъяв;н/б;3;ед;3" reg="подобаетъ" src="ПОДОБАЕТЪ">подобаетъ</w>
209       <w xml:id="DmPrlc.153" ana="мест;м;вин;мн;ср" lemma="иже" reg="яже" src="ЯЖЕ">же</w>
210       <w xml:id="DmPrlc.154" ana="гл;изъяв;аор гл;1;мн" lemma="вид+ти" reg="вид+хомъ"
          ↪  src="ВИД+ХОМЪ">видѣхомъ</w>
211       <pc xml:id="DmPrlc.155">.</pc>
212       <lb n="5"/>
213       <w xml:id="DmPrlc.156" ana="союз" lemma="и" reg="и" src="И">и</w>
214       <w xml:id="DmPrlc.157" ana="союз" lemma="аще" reg="аще" src="АЩЕ">аще</w>
215       <w xml:id="DmPrlc.158" ana="союз" lemma="и" reg="и" src="И">и</w>
216       <w xml:id="DmPrlc.159" ana="прил;м;вин;мн;ср" lemma="прежнии" reg="прежняя" src="ПРЕЖНRА">прежнѧа</w>
217       <w xml:id="DmPrlc.160" ana="союз" lemma="и" reg="и" src="И">и</w>
218       <w xml:id="DmPrlc.161" ana="гл;изъяв;аор гл;1;мн" lemma="слышати" reg="слышахомъ"
          ↪  src="СЛЫШАХОМЪ">слышахомъ</w>
219       <pc xml:id="DmPrlc.162">.</pc>
220       <w xml:id="DmPrlc.163" ana="сущ;jo;дат;ед;ср" lemma="писание" reg="писанию" src="ПИ&amp;САНIЮ">пи<lb
          ↪  n="6"/>санїю</w>
221       <w xml:id="DmPrlc.164" ana="прил;тв;вин;мн;ср" lemma="истинныи" reg="истинная"
          ↪  src="ИСТИННАА">истиннаа</w>
222       <w xml:id="DmPrlc.165" ana="инф" lemma="положити" reg="положити" src="ПОЛОЖИТИ">положити</w>
223       <pc xml:id="DmPrlc.166">.</pc>
224     </ab></body></text>
225   </TEI>
```