



www.volsu.ru

ГЛАВНАЯ ТЕМА НОМЕРА

DOI: <https://doi.org/10.15688/jvolsu2.2021.6.1>

UDC 811.161.1'04
LBC 81.411.2-03



Submitted: 07.07.2021
Accepted: 30.09.2021

PARALLEL CORPUS OF *THE SLAVONIC PARIMEJNIK* COPIES: COMPOSITION, STRUCTURE, ANALYTICAL MARKUP¹

Victor A. Baranov

Kalashnikov Izhevsk State Technical University, Izhevsk, Russia

Abstract. The article gives grounds for marking up machine-readable transcriptions of medieval Slavic manuscripts, which serve as textual material for the historical corpora, the excerpts are viewed as possessing valuable codicological or textological characteristics. Composition analysis of four manuscripts from *the Slavonic Parimejnik* (12th – 14th cc.) and modeling their structure with application of generally accepted tools of linguistic analysis have enabled solving the following tasks: analytical units identification, text format elaboration and search process algorithmization based on the natural language units characteristics. A suggested format of lectionaries description includes data on the section (sub-section) of the liturgical year, the number of the lectionary in this section, textual composition in relation to the texts of the Bible, the topic. The format of dates, days and time of lectionaries reading throughout the year includes data on the section (sub-section), the date of the fixed calendar or the week and the day within the section (sub-section) of the nonfixed calendar, the time of the church service, the event to which the service is devoted. The markup of the Parimejnik texts included indication of the excerpts boundaries and determination of their relations with the dictionary unit with the aim to ensure the parallel corpus of four Slavic manuscripts of the Parimejnik with ranging the corresponding excerpts at the level of the lectionaries and Bible verses within the corpus “Manuscript” (manuscripts.ru).

Key words: Medieval Slavonic manuscripts, Parimejnik (Old Testament lectionary), structure and composition, markup, parallel corpus.

Citation. Baranov V.A. Parallel Corpus of *the Slavonic Parimejnik* Copies: Composition, Structure, Analytical Markup. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], 2021, vol. 20, no. 6, pp. 6-17. (in Russian). DOI: <https://doi.org/10.15688/jvolsu2.2021.6.1>

УДК 811.161.1'04
ББК 81.411.2-03

Дата поступления статьи: 07.07.2021
Дата принятия статьи: 30.09.2021

ПАРАЛЛЕЛЬНЫЙ КОРПУС СПИСКОВ СЛАВЯНСКОГО ПАРИМЕЙНИКА: СОСТАВ, СТРУКТУРА, АНАЛИТИЧЕСКАЯ РАЗМЕТКА¹

Виктор Аркадьевич Баранов

Ижевский государственный технический университет им. М.Т. Калашникова, г. Ижевск, Россия

Аннотация. Обоснована необходимость разметки машиночитаемых транскрипций средневековых славянских рукописей, являющихся текстовым материалом исторических корпусов, на уровне аналитических фрагментов – частей, обладающих кодикологически или текстологически значимыми характеристиками. На примере анализа состава четырех рукописей славянского паримейника (XII–XIV вв.) и разра-

ботки его структурной модели продемонстрировано решение задач идентификации аналитических единиц, сохранения традиций их описания, создания условных форматов, обеспечения поиска и представления их характеристик с помощью естественного языка. Предложен формат описания паримий, включающий сведения об отделе (подотделе) церковного года, номере паримии в этом отделе, текстовом составе в его отношении с текстами Библии, теме. Разработан формат описания дат, дней и времени чтения паримий в течение года, который включает сведения об отделе (подотделе), дате по неподвижному или неделе и дне в пределах отдела (подотдела) по подвижному календарю, о времени церковной службы, о событии, которому посвящена служба. Разметка списков паримейника (указание границ каждого фрагмента, установление связи фрагмента со словарной единицей) позволяет демонстрировать в корпусе «Манускрипт» (manuscripts.ru) параллельный корпус четырех славянских списков паримейника, выровненных по паримиям и стихам Библии.

Ключевые слова: средневековые славянские рукописи, паримейник, структура и состав, разметка, параллельный корпус.

Цитирование. Баранов В. А. Параллельный корпус списков славянского паримейника: состав, структура, аналитическая разметка // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2021. – Т. 20, № 6. – С. 6–17. – DOI: <https://doi.org/10.15688/jvolsu2.2021.6.1>

Введение

Исследователи исторических текстовых источников всегда исходили из понимания ограниченности анализируемого материала, обусловленной количеством рукописей, которые сохранились и доступны для анализа. Поэтому отношение текстологов и лингвистов к данным средневековых рукописей и текстов особое: важны все факты изучаемого явления в их соотношении, в том числе и количественном, друг с другом и в противопоставлении фактам альтернативного способа выражения или альтернативного значения. Это позволяет говорить о том, что традиционный исторический метод лингвистического анализа средневековых рукописных источников использует по сути корпусные методики извлечения и систематизации лингвистических данных [Баранов, 2015], а корпусное понимание текста как единственно доступного нам объекта наблюдения и изучения имеет ту же методологическую основу, что и база историко-лингвистических исследований, опирающаяся на факты дошедших до нас письменных источников.

Сегодня корпусные методы, использующие в первую очередь автоматизированные способы и приемы извлечения, количественного сопоставления и статистической оценки лингвистических форм, конструкций и их значений, активно применяются не только для решения прикладных задач поиска и анализа текстов (контекстов), их генерации, создания систем взаимодействия человека и машины на естественном языке, но и для научных це-

лей – оценки текстов с точки зрения близости и контрастности, подтверждения лингвистических гипотез, нахождения в текстах неизвестных закономерностей и многих других.

До недавнего времени применение корпусных методов в работе со средневековыми славянскими рукописями было невозможным: отсутствовали машиночитаемые текстовые ресурсы достаточного объема. Но за последние десять – пятнадцать лет ситуация несколько (пока, наверно, рано использовать слова *существенно, принципиально, кардинально*) изменилась: у отдельных исследователей и коллективов появились достаточно большие коллекции наборных копий (транскрипций) средневековых текстов (рукописей), которые позволяют или применить к ним простейшие автоматизированные приемы выборки, счета, сортировки, предоставляемые, в частности, текстовыми процессорами и электронными таблицами, или осуществить их обработку с помощью специализированных программ, например AntConc, Voyant, Gephy и др., для подготовки конкордансов, получения количественных и статистических сведений, визуализации текстовых закономерностей, или, создав на основе транскрипций корпус, использовать возможности корпусных менеджеров для поиска, выборки, демонстрации и анализа лингвистической информации².

Как известно, анализ лингвистических единиц включает в себя в том числе и этап их сопоставления друг с другом по каким-либо формальным и содержательным признакам: от текстовых (время, автор, стиль и жанр) до

собственно речевых и языковых (позиция, грамматическая форма, семантика). Поэтому основой любой коллекции и корпуса является разметка – метаразметка текстов, частеречная и морфологическая разметка лингвистических единиц, синтаксическая (их грамматических связей), тематическая, семантическая (отношений слов с другими словами) и др. Наличие разметки позволяет не только найти в большой коллекции единицы с необходимыми значениями, но и сформировать выборки, единицы которых обладают наборами идентичных (аналогичных) или противопоставленных форм или значений.

В настоящее время для любой коллекции или корпуса базовыми являются мета- и грамматическая разметки. Дополнительные зависят от направленности проекта, исследовательских задач, которые решает коллектив, от наличия инструментария, позволяющего в автоматическом или автоматизированном режиме осуществлять разметку, и, конечно, от самого текстового материала.

Если для современных корпусов с их огромным количеством данных значимые наблюдения успешно осуществляются на основе временных, жанровых, авторских, гендерных, возрастных и под. противопоставлений, то для существенно меньших по объему средневековых коллекций при сохранении значимости временных и жанровых характеристик особое внимание при анализе речевых и языковых особенностей уделяется свойствам отдельных частей рукописей и текстов. Понятно, например, что при выборке данных из кодекса, переписанного несколькими писцами, необходимо предоставить пользователю возможность получить сведения о каждой из частей отдельно³, при наличии в рукописи нескольких текстов анализ должен осуществляться с учетом того, что лингвистические единицы принадлежат разным текстам⁴, а следовательно, их границы должны быть указаны. При компилятивном характере текста информация о текстологически важных фрагментах должна быть внесена в машиночитаемую копию⁵. Таким образом, учет текстологических, кодикологических и лингвистических особенностей средневековых письменных памятников при их исследовании с помощью корпусных методов должен быть обеспечен соответствующей разметкой. И чем

более детальной она будет, тем больше возможностей для сопоставления материала будет предоставлено лингвисту.

Данная работа посвящена демонстрации способов решения прикладной текстологической задачи – представления структуры и состава средневекового славянского паримейника в виде модели и в виде разметки четырех славянских списков XII–XIV вв. в базе данных исторического корпуса «Манускрипт»⁶.

Наличие в базе данных корпуса структурированных сведений о фрагментах списков позволяет формировать выборки лингвистического материала с учетом значений их частей: а) создавать подкорпусы из фрагментов одной рукописи с разными аналитическими характеристиками⁷ и подкорпусы из фрагментов разных списков с идентичными значениями; б) сопоставлять подкорпусы с помощью инструментов, анализирующих количественные характеристики их лингвистических единиц; в) представлять на экране соответствующие друг другу части рукописей в виде параллельного корпуса списков одного текста.

Материал и методы

Рукописи паримейника в корпусе

Как известно, паримейник (в греческой традиции – профитологий) – сборник избранных чтений в основном из Ветхого Завета. Паримии читаются в дни Великого поста, на Страстной неделе, в дни праздников и памятных событий [Алексеев, 2008, с. 161–162]. Паримии содержат тексты, в той или иной степени соотносимые с поминаемым в этот день событием [Алексеев, 2008, с. 163 и др.] и сгруппированные в три отдела – Рождественско-Богоявленский, Триодный и Месяцесловный, а также указания на чтения, исполняющиеся во время службы до, между или после паримий.

В настоящее время в корпусе «Манускрипт» имеется четыре списка паримейника: Лазаревский (Сковородский), XII в. (РГАДА 1988, с. 119–121; СК 2002, с. 641–643), Захарьинский 1271 г. (СК 1984, с. 205–206), Федоровский, XIII в. (СК 1984, с. 310–311) и Троицкий, XIV в. (ПС 1966, с. 229; ОР 1878, с. 3) (см. список источников; далее – Лаз, Зах, Фед, Тр соответственно)⁸.

Рукописи в разной степени сохранности. Наименьшее количество утрат в Зах: утраты между 15 и 16 листами (нет одной паримии на Богоявление и конца предыдущей паримии и начала следующей) и между 195 и 196 листами (отсутствует несколько паримий Страстной пятницы и Страстной субботы) (СК 1984, с. 206). Наибольшее – в Фед: большие утраты во всех отделах – нет нескольких Рождественских паримий, паримий на Богоявление и Водосвятие, отсутствуют паримии первой недели поста и понедельника и вторника второй недели и мн. др. (СК 1984, с. 311).

Две рукописи Лаз и Зах созданы несколькими писцами. Текст Зах переписан попом церкви Св. Дмитрия в Новгороде Захарией, его сыном Олуферием и еще двумя писцами (СК 1984, с. 205–206), текст Лаз – тремя, вторым из которых был поп новгородского монастыря Св. Лазаря Домка, переписавший также большую часть Милятина евангелия (РНБ, Ф.п.1.7) [Михеев, 2019, с. 27–37; Мольков, 2020, с. 35; и др.].

Установление состава и структуры рукописей

Паримии и их идентификация

В печатных изданиях паримейников традиционно принято в качестве идентифицирующего признака паримии указывать ее номер. Так, в [Monumenta..., 1939, р. 600–604; 1980–1981, р. 163–170] дана сквозная нумерация паримий в 71-й греческой рукописи IX–XIV вв. (сведения о количестве по: [Алексеев, 2008, с. 160]) от L1a до L76c (L – лат. lectio ‘чтение’). Один и тот же номер арабскими цифрами присвоен нескольким паримиям, читающимся в течение одной службы или на нескольких службах одного дня: восемь паримий навечерия Рождества имеют номер L1, две паримии, читающиеся в Сырную среду, – номер L3 и т. д.; литерная нумерация используется для идентификации паримий внутри службы или дня: восемь паримий Рождества последовательно пронумерованы от L1a до L1h, паримия на 6-м часе имеет номер L3a, на вечерне – L3b и т. д.

Несколько иным способом паримии идентифицированы в обобщающем труде [Алексеев, 2008, с. 234–245]. А.А. Алексеев

использует различные приемы: а) указание на службу определенного дня неподвижного или подвижного календаря или комментарий; б) нумерация; в) указание на соотношение с чтениями Библии; г) тема паримии. В перечне паримий указываются отделы (Рождественско-Богоявленский, Триодный, Месяцесловный), подотделы (например, Великий пост, Лазарева суббота и Страстная неделя), день (например, Страстной понедельник, Страстной вторник и др.) и служба (например, на вечерне, на всенощном бдении, на 6-м часе и др.). Паримии на Рождество, Богоявление, Водосвятие и на вечерне Страстной субботы нумеруются по порядку следования. В другие дни идентификация осуществляется с помощью указания на стихи Библии и темы. См., например, три паримии на Вознесение:

Ис. 2.2-11 Суд Господень

Ис. 62.10-63.3-9 Не ангел, но сам Господь спас их

Зах. 14.1-11 Живые воды потекут из Иерусалима.

А также с помощью комментария о службе паримии в 1-й Понедельник Великопоста:

Ис. 1.1-20 Омойтесь, очиститесь

Быт. 1.1-13 Три дня творения

Притч. 1.1-20 Начало премудрости – страх Господень («Прим. 1. В течение шести недель Великопоста Ис читается на 6-м часе, Быт и Притч – на вечерне» [Алексеев, 2008, с. 235]).

Удобство представления состава паримейников в обоих случаях бесспорно: в первом случае каждая паримия имеет уникальный номер, позволяющий соотносить ее со службой и местом в службе, во втором – паримия имеет несколько характеристик, позволяющих установить в том числе и их значение.

В то же время и тот, и другой способ обладают недостатками с точки зрения идентификации паримий: в первом – нет указания на то, что некоторые паримии читаются несколько раз в году (для этого используются дополнительные таблицы соответствий), во втором – отсутствует единый способ обозначения паримий.

Разметка машиночитаемой транскрипции предполагает указание объекта (в данном случае – границ фрагмента) и приписывание ему значений, которые могут быть как уникальными, альтернативными, так и неуникальными, совпадающими со значениями других объектов идентичного типа. В любом случае необходим перечень параметров, значения которых позволяют: а) однозначно указать каждый из объектов; б) описать его существенные характеристики. При разработке компьютерных систем еще одно условие должно быть учтено – необходимость автоматической обработки значений, что возможно при наличии формализованных значений объектов.

Понятно, что при разметке паримий желательно сохранить связь с печатной традицией: для соотнесения пользователем результатов работы в корпусе с имеющейся в литературе информацией о составе паримейников необходимо присвоить фрагментам значения, позволяющие осуществлять поиск и формирование подкорпусов на основе идентичных и альтернативных значений, и обеспечить возможность автоматического упорядочения при выводе на экран.

В настоящее время для разметки паримейников в базе данных «Манускрипт» используются следующие параметры: номер паримии, сведения о соотношении текста с текстом Библии, тема паримии, дата и время ее чтения.

Основным параметром является номер, включающий по традиции указание на тип фрагмента (Р – паримия), номер отдела / подотдела / дня (цифровая нумерация), порядок следования в течение дня (литерная нумерация), альтернативные паримии (дополнительная литерная нумерация), а кроме того, номер другого отдела (как отсылка к той же паримии, читающейся на другой службе):

Р7а = Р42аа,

где Р – паримия, 7 – 1-я Среда Великого поста, а – первая паримия Среды, = – паримия читается также на другой службе, 42 – служба на праздник Преполовления, аа – паримия, альтернативная паримии 42а.

Дополнительные параметры соотносят каждую паримию с текстами Библии и содержанием перикопы, которые даны в [Алексеев, 2008, с. 234–245]:

Р7а = Р42аа Ис. 2.3-11 «Перекуют мечи на орала»,

Р42аа = Р7а Ис. 2.3-11 «Перекуют мечи на орала».

В теме, вслед за А.А. Алексеевым, в квадратных скобках «приводится содержание некоторых опущенных пассажей книг Бытия и Притч» [Алексеев, 2008, с. 236]: Р13b Быт. 5.1-24 «Родословие Адама [Рождение Ноя]», а в случае различных формулировок темы одной паримии – обе формулировки: Р2i = Р37b Исх. 2.5–10 «Обретение Моисея на Ниле = Дочь фараона находит Моисея».

Имеются случаи, требующие уточнения предложенной системы идентификации паримий: а) вариативность текстов паримий; б) наличие в рукописи нескольких паримий, соответствующих по тексту одной в другом списке.

В первом случае различаем:

а) содержательную вариативность, при которой (частично) меняется тема текста, ср.: Р1а = Р2а = Р5b Быт. 1.1–13 «Три дня творения» и Р41b Быт. 1.1–5 «Первый день творения». Соответственно, паримии квалифицируются как разные;

б) формальную вариативность, при которой содержание не меняется, ср. отсутствие контекста, соответствующего стиху Пр. 5.7, в списке Тр (20b–21b) в паримии Р11с Пр. 5.1–15 «Не приближайся к дверям блудницы», или наличие дополнительных контекстов в конце паримии Р14с Пр. 6.20 – 7.1а «Опасность прелюбодеяния [Опасность от блудницы]» в Зах (68а – 69а) и др. Соответственно, признается наличие в списках одной и той же паримии;

в) отсылочную вариативность, при которой в рукописи выписывается лишь заголовок и начало паримии, а также дается отсылка к полному тексту. См. 3-ю паримию на Воздвижение креста (14 сентября) Р51с Ис. 60.11–16 «Будешь насыщаться молоком народов», в которой приводится только начало 11-го стиха и отсылается ко второй части второй паримии на вечерне Страстной субботы Р41с Ис. 60.1–16 «Светися, / Иерусалиме, слава Господня на тебе возсияла»:

чтѣ · ꙗ · ѿ пррѣ ісѣі / тако глѣть ꙗ ѿверста / будутъ
врата твоѣ ꙗꙗме · ꙗ ꙗꙗ · въ вели ісѣ · вечеръ · втораго
чтѣ · / ѿ полу · (Тр 116d)

22 марта и 25 апреля по Юлианскому календарю. Аналогичны и различия в пределах суток: ср. 24-часовой отсчет современных суток, которые начинаются в полночь, и порядок следования служб в течение суток церковного календаря, в котором новый день начинается вечером, до полуночи, каждая служба имеет свое наименование (1-й час, 3-й час и т. д., вечерня, повечерие, утреня и др.) и не привязана точно к конкретному часу (например, служба 1-го часа начинается около 7 часов утра, вечеря – около 9 часов вечера и др.), а первой службой нового дня является вечерня.

Все сказанное позволяет сделать вывод о том, что разметка богослужебного текста с указанием дня и времени чтения не может быть единообразной: разметка должна содержать сведения о двух годовых циклах – неподвижном и подвижном.

Согласно действующему стандарту представления дат и времени [ГОСТ Р 7.0.64-2018], идентификация дней может быть сделана на основе как календарных, так порядковых дат. В первом случае используется порядковый номер месяца в году и порядковый номер дня в календарном месяце, во втором – может быть использован порядковый номер недели и порядковый номер дня в неделе.

Расширенный формат календарной даты, предписываемый стандартом, – YYYY-MM-DD, порядковой даты – YYYY-Www-D, где YYYY – номер года, MM – номер месяца в году, DD – номер дня в месяце, Www – порядковый номер недели, D – номер дня в неделе (1 – понедельник и т. д.). Расширенный формат времени – T_{hh}:mm:ss, где T – метка области времени: часы, минуты, секунды. Стандарт предусматривает возможность модификаций форматов по соглашению сторон, обменивающихся информацией.

Современный стандарт может быть использован для разметки текстов паримейников (шире – любых богослужебных текстов) с некоторыми дополнениями и соглашениями. Так, для разметки средневековых славянских текстов целесообразно: а) использовать Юлианский календарь современного православного богослужения (при необходимости даты могут быть переведены в даты современного Григорианского календаря по новому сти-

лю); б) учитывать разделение Пасхального года на отделы, подотделы (циклы, подциклы); в) порядковый номер недели исчислять от первой недели каждого из отделов и подотделов; г) начало служб условно соотнести с тем или иным временем (часом) суток; д) службы, посвященные одному событию, размечать как службы одного определенного дня (вечерняя служба Рождества Христова, совершаемая вечером 24 декабря, в связи с тем, что новый день начинается до полуночи, должна получить дату 25 декабря); е) использовать современную терминологию для обозначения церковных понятий *седмица* и *неделя*, а именно: *неделя* и *воскресенье*. Эти соглашения позволяют использовать форматы стандарта и его терминологию, иметь уникальное значение текстов с точки зрения времени их чтения в течение года, увеличивать количество уникальных значений в случае необходимости (например, использовать аналогичную разметку для других служебных текстов), осуществлять поиск текстов на основе их временных значений и др.

Проиллюстрируем сказанное примерами. Значения даты и времени:

1) текстов, читающихся на службе Рождества Христова: **Ch|--12-25T21|**-, где обозначение отдела (цикла) – **Ch** = Christmas = Рождество; указание на дату по неподвижному календарю – **|--12-25** = любой год-декабрь-25; сведения о службе – **T21** = вечерня; указание на день по подвижному календарю – **|** = значение отсутствует;

2) текстов, читающихся на одной из служб Великого поста: **GL|--W01-1T12**, где **GL** = Great Lent = Великий пост, **|** – дата по неподвижному календарю отсутствует, **|--01-1** – понедельник 1-й недели, **T12** – служба 6-го часа.

Из примеров видно, что формат имеет три части:

Div|--MM-DDT_{hh}|--Www-dT_{hh},

где Div – отдел или подотдел: Рождество (Ch – Christmas), Богоявление (Ep – Epiphany), Водосвятие (BoW – Blessing of Water), Недели перед Великим постом (PWL – Preparatory Weeks for Lent), Великий пост (GL – Great Lent), Страстная неделя (HW – Holy Week), Пентекостарий (Pnt – Pentecost), Недели после Пятидесятницы (WaP – Weeks after Pentecost), Месяцеслов (Mnl – Menologium);

|--MM-DDThh – формат даты и службы неподвижного календаря;

|--Www-dThh – формат дня и службы подвижного календаря, где Www – аббревиатура и номер недели от начала отдела, d – номер дня недели, Thh – идентификатор службы: 21 – вечерня, 22 – повечерие, 23 – всенощная, 24 – полунощница, 02 – утренняя, 07 – 1-й час, 09 – 3-й час, 12 – 6-й час, 13 – литургия, 15 – 9-й час.

Таким образом, каждая служба кодируется уникальным идентификатором, присваиваемым словарной единице словаря фрагментов по времени чтения, с которыми связываются соответствующие друг другу фрагменты разных списков.

Важным условием, также определяющим точность и функциональность разметки, является необходимость указания не только дат и времени служб, но и событий, которым они посвящены. Предложенный комбинированный формат содержит сведения о некоторых праздниках, Страстной неделе, некоторых памятных событиях, но он является условной машиночитаемой аббревиатурой и не позволяет без специальной обработки предоставить пользователю возможность осуществлять поиск данных на естественном языке и выводить на экран необходимые для понимания фрагментов сведения.

В настоящее время каждая единица словаря, обеспечивающего разметку рукописей в соответствии с датой и временем чтения фрагмента, снабжается кратким комментарием, содержащим словесную информацию об отделе (подотделе), дате, времени и событии:

Ch|--12-25T21|-, комментарий: Рождество. 25 декабря. Вечерня;

GL|--W01-1T12, комментарий: Великий пост. 1-я неделя. Понедельник. 6-й час;

GL|--W06-6T23, комментарий: Великий пост. 6-я неделя. Лазарева суббота. Всенощное бдение; Pnt|--W06-4T21, комментарий: Пентекостарий. Вознесение. 6-я неделя. Четверг. Вечерня;

WaP|--W01-7T21, комментарий: Недели после Пятидесятницы. 1-я неделя. Воскресенье. Неделя Всех Святых. Вечерня;

Mn|--09-08T21|-, комментарий: Месяцеслов. 8 сентября. Рождество Богородицы. Вечерня и др.

Дублирование информации в двух способах идентификации каждой словарной едини-

цы позволяет обеспечить на данном этапе поиск и демонстрацию фрагментов без создания специальных процедур перевода одного обозначения в другой. Кроме того, немаловажным является удобство поиска и редактирования словарных единиц при наличии значений в словесной форме.

Результаты и обсуждение

Применение корпусных методов для изучения славянских письменных памятников невозможно без подготовки исторических корпусов, тремя базовыми составляющими которых являются: машиночитаемые тексты, их разметка и средства обработки, поиска, демонстрации и анализа лингвистических данных.

И подготовка транскрипций, максимально близких к оригиналу, и создание лингвистической разметки предполагают создание необходимых форматов, средств и инструментов, которые позволяют передать графические, орфографические, грамматические составляющие средневекового текста в соответствии с его особенностями.

Необходимость исследовать рукописи как объекты, состоящие из частей с разными кодикологическими (например, разные писцы) и текстологическими (например, тексты разных жанров и назначения) характеристиками, без учета которых анализ лингвистических данных будет усредненным, неточным, заставляет решать задачи моделирования и разметки аналитической структуры рукописей и текстов.

Разметка списков паримейника как текстовых структур, состоящих из фрагментов с различными характеристиками, требует решения нескольких задач:

а) идентификация каждой аналитической единицы;

б) сохранение связи традиций их описания с машиночитаемым представлением характеристик;

в) создание идентифицирующих форматов;

г) обеспечение поиска и демонстрации характеристик фрагментов с помощью естественного языка.

Предложенные решения разметки паримей, их составных частей, комментариев, а

также дат и времени чтения фрагментов обеспечивают идентификацию каждой из аналитических единиц с помощью нескольких параметров: границ в пределах списка, связью с единицами словаря, идентифицирующего формализованного значения, дополнительных характеристик.

Так, значения единицы **P9c = P50b** Пр. 3.19–34 «Смирением дает благодать» указывают на ее место в традиционно принятом в паримейнике ряду паримий – это третья (**c**) паримия (**P**) службы в пятницу 1-й недели Великого поста (**9**), на ее текстовый состав, соответствующий 19–34-му стихам 3-й главы книги Притч, на ее тему и на то, что эта же паримия является вторым чтением (**b**) на службе 13 сентября (**50**).

При решении задачи идентификации фрагментов, читающихся в различные дни года и на различных службах: а) предложены решения, сохраняющие традицию; б) использованы действующие стандартные соглашения, определяющие даты и время событий; в) разработан машиночитаемый формат, совмещающий неподвижный и подвижный календари; г) применяются значения словарных единиц, обеспечивающие поиск и демонстрацию данных, понятные пользователю.

Так, идентификационное значение словарной единицы словаря дат и времени **GL|–W01-5T21** указывает на исполнение текста на вечерне (**T21**) в пятницу (**-5**) первой недели (**-W01**) Великого поста (**GL**), а значение **Mnl|–09-08T21|–** – на чтение текста на вечерне (**T21**) 8-го (**-08**) сентября (**-09**) по неподвижному календарю (**Mnl**).

Выводы

Анализ сложных по структуре и составу текстов, разработка модели разметки данных в соответствии с особенностями текстов, а также с требованиями, которые предъявляются к формализованным описаниям, обеспечивающим обработку, поиск и демонстрацию данных, являются необходимым этапом при создании глубоко размеченных машиночитаемых текстовых ресурсов на основе средневековых славянских рукописей.

Аналитическая разметка позволяет расширить условия подготовки запросов и учиты-

вать при поиске лингвистических данных в том числе и характеристики кодикологически и текстологически важных фрагментов, что обеспечивает дифференцированный анализ речевых форм и значений в пределах рукописи.

Значения единиц, выделенных при разметке, дают возможность осуществлять поиск соответствующих друг другу фрагментов различных списков, создавать выборки, обладающие идентичными или противопоставленными значениями, упорядочивать единицы выборок, выводить результаты на экран.

Наличие у аналитических фрагментов соответствующих друг другу значений позволяет выровнять их при демонстрации и предоставить пользователю возможность одновременного просмотра на уровне наиболее крупных (дата, день и время), крупных (паримии) и небольших (стихи) фрагментов, возможность сопоставления их места в списках, обнаружения лингвистических разночтений.

Вместе с подготовкой транскрипции и разработки процедур и программ поиска и демонстрации данных описание частей рукописей и аналитическая разметка являются составными этапами создания параллельного корпуса.

ПРИМЕЧАНИЯ

¹ Исследование выполнено при финансовой поддержке РФФИ, проект № 20-512-18001 «Средневековые тексты в современном контексте (новые методы и принципы представления средневековых текстов сегодняшним пользователям)».

The reported study was funded by RFBR, project no. 20-512-18001 “Medieval Texts in the Contemporary Context (Modern Methods and Principles of Presentation of Medieval Texts to the Nowadays Users)”.

² См., например, функциональные возможности исторических корпусов НКРЯ (ruscorpora.ru), диакронического корпуса Чешского национального корпуса (<https://www.korpus.cz/>), корпуса «Великие миеи-четыре» (<http://www.vmc.uni-freiburg.de/Mens/>), проекта «The Russian Diachronic Online Corpus» (<https://www.slawistik.hu-berlin.de/de/member/meyerrol/subjekte/trudi>), корпуса «The TOROT Treebank» (<https://torottreebank.github.io/>), Санкт-Петербургского корпуса агиографических текстов (<http://project.phil.spbu.ru/scat/page.php?page=project>), исторического корпуса «Манускрипт» (manuscript.ru) и некоторых других.

³ См., например, работу [Гиппиус, 1989], в которой дан анализ причастных форм, местоимений и имен в разных почерках Захарьинского паримейника.

⁴ См., например, лингвостатистический анализ анонимных текстов Толстовского сборника на фоне авторских в [Zholobov, Baranov, Novak, 2020; Баранов, Жолобов, 2021].

⁵ См. постановку задачи в: [Baranov, Zuga, 2021].

⁶ Данная работа является продолжением публикации [Баранов, 2021], в которой показаны результаты подготовительных работ по созданию параллельного корпуса списков славянского паримейника.

⁷ Под аналитическими характеристиками понимаются характеристики составных частей текста и/или рукописи, обладающие кодикологически и текстологически значимыми особенностями. Под аналитической разметкой – указание границ составных частей текста и/или рукописи и их индивидуальных или совпадающих с другими частями значений.

⁸ Машиночитаемые копии рукописей подготовлены в разное время А.И. Кузовенковой, А.О. Штатновой (Казанский (Приволжский) федеральный университет), Г.А. Мольковым (Институт лингвистических исследований РАН, Институт русского языка им. В.В. Виноградова РАН), загрузка в корпус «Манускрипт» осуществлена В.А. Барановым (Ижевский государственный технический университет им. М.Т. Калашникова) и Р.М. Гнутиковым (Удмуртский государственный университет), сверка с оригиналом – О.Ф. Жолобовым (Казанский (Приволжский) федеральный университет) и М.О. Новак (Институт русского языка им. В.В. Виноградова РАН), корректура транскрипций – М.О. Новак и В.А. Барановым (см.: <http://manuscripts.ru/mns/portal.main?p1=54>).

СПИСОК ЛИТЕРАТУРЫ

- Алексеев А. А., 2008. Библия в богослужении. Византийско-славянский лекционарий. СПб. : Нестор-История. 268 с.
- Баранов В. А., 2015. Исторический корпус как цель и инструмент корпусной палеославики // *Scripta & e-Scripta*. Vol. 14/15. P. 39–62.
- Баранов В. А., 2021. Параллельный корпус славянских списков паримейника: материал и постановка задачи // *Корпусная лингвистика – 2021*. (В печати).
- Баранов В. А., Жолобов О. Ф., 2021. Лингвостатистический анализ древнерусской антологии (РНБ, Ф.п.1. 39) // *Социально-экономическое управление: теория и практика*. № 2 (45). С. 50–60. DOI: 10.22213/2618-9763-2021-2-50-60.

Гиппиус А. А., 1989. Система формальных признаков языка древнерусской письменности как предмет лингвистического изучения // *Вопросы языкознания*. № 2. С. 93–110.

ГОСТ Р 7.0.64-2018. Представление дат и времени. Общие требования. М. : Стандартинформ. 40 с. URL: <http://docs.cntd.ru/document/1200159341> (дата обращения: 01.06.2021).

Михеев С. М., 2019. Миней двух Домок: еще раз о писцах служебных миней из новгородского Лазарева монастыря // *Слов'яне = Slovĕne*. Vol. 8, № 2. С. 7–56. DOI: 10.31168/2305-6754.2019.8.2.1.

Мольков Г. А., 2020. Формирование орфографических систем в древнерусской письменности XI – начала XIII века : дис. ... д-ра филол. наук. СПб. 494 с. URL: <https://iling.spb.ru/dissovet/theses/molkov/thesis.pdf> (дата обращения: 01.06.2021).

Baranov V. A., Zuga O. V., 2021. Correlation and Cluster Analysis of Fragments of the Earliest Slavonic Gospels // *El'Manuscript 2021. 8th International Conference on Textual Heritage and Information Technologies (Freiburg im Breisgau, Germany, 12.04.2021–15.04.2021)* : booklet of abstracts / ed. by J. Besters-Dilger, A. Rabus. Freiburg im Breisgau : Albert-Ludwigs-Universität Freiburg. P. 11–12.

Monumenta Musicae Byzantinae. Lectionaria. Vol. I. Prophetologium. Pars. I. Lectiones Nativitatis et Epiphaniae, 1939 / Edd. C. Höeg, G. Zuntz. Copenhagen : Hauniae : Levin & Munksgaard. 612 p.

Monumenta Musicae Byzantinae. Lectionaria Vol. I. Prophetologium. Pars. II. Lectiones anni Immobilis, 1980–1981 / Edd. G. Engberg. Copenhagen : Hauniae : Levin & Munksgaard. 314 p.

Zholobov O. F., Baranov V. A., Novak M. O., 2020. Anonymous vs. Attributed: Cluster Analysis of Tolstovskii Sbornik Texts and Its Interpretation in Terms of Cultural Heritage // *Journal of Siberian Federal University. Humanities & Social Sciences*. 13 (12). P. 2056–2081. DOI: 10.17516/1997-1370-0704.

ИСТОЧНИКИ

- Зах* – Захарьинский паримейник 1271 г. // РНБ. Ф.п.1.13. 264 л.
- Лаз* – Лазаревский (Сковородский) паримейник, XII в. // РГАДА. Ф. 381. Оп. 1. Тип. 50. 126 л.
- Тр* – Троицкий паримейник, XIV в. // РГБ. Ф. 304/1. № 4. 142 л.
- Фед* – Федоровский II паримейник, XIII в. // РГАДА. Ф. 381. Оп. 1. Тип. 60. 107 л.

**КАТАЛОГИ И ОПИСАНИЯ
РУКОПИСНЫХ СОБРАНИЙ**

- ОР 1878* – Описание славянских рукописей библиотеки Свято-Троицкой Сергиевой Лавры. Ч. 1. Священное писание. №4 (2005). М.: [б. и.], 1878. 830 с. URL: <https://www.prlib.ru/item/434785> (дата обращения: 01.06.2021).
- ПС 1966* – Предварительный список славяно-русских рукописей XI–XIV вв., хранящихся в СССР (для «Сводного каталога рукописей, хранящихся в СССР, до конца XIV в. включительно») // Археографический ежегодник за 1965 г. М.: [б. и.], 1966. С. 177–273.
- РГАДА 1988* – Каталог славяно-русских рукописных книг XI–XIV вв., хранящихся в ЦГАДА СССР. Ч. 1 / О. А. Князевская [и др.]. М.: [б. и.], 1988. 123 с.
- СК 1984* – Сводный каталог славяно-русских рукописных книг, хранящихся в СССР. XI–XIII вв. М.: Наука, 1984. 406 с.
- СК 2002* – Сводный каталог славяно-русских рукописных книг, хранящихся в России, странах СНГ и Балтии. XIV век. Вып. 1. Апокалипсис – Летопись Лаврентьевская / редкол.: О. А. Князевская [и др.]. М.: Индрик, 2002. 768 с.

REFERENCES

- Alekseev A.A., 2008. *Bibliya v bogoslužhenii. Vizantiysko-slavyanskiy leksionariy* [Bible in Worship. Byzantine-Slavic Lectionary]. Saint Petersburg, Nestor-Istoriya Publ. 268 p.
- Baranov V.A., 2015. Istoricheskiy korpus kak tsel i instrument korpusnoy paleoslavistiki [The Historical Corpus As the Target and Tool of Corpus Paleoslavistics]. *Scripta & e-Scripta*, vol. 14/15, pp. 39–62.
- Baranov V.A., 2021. Parallelnyy korpus slavyanskikh spiskov parimeynika: material i postanovka zadachi [Parallel Corpus of Slavonic Parimejnik Manuscripts: Material and Problem Statement]. *Korpusnaya lingvistika – 2021*. (In press).
- Baranov V.A., Zholobov O.F., 2021. Lingvostatisticheskiy analiz drevnerusskoy antologii (RNB, F.p.I. 39) [Lingvo-Statistical Analysis of the Old Russian Anthology (RNL, F.p.I. 39)]. *Sotsialno-ekonomicheskoe upravlenie: teoriya i praktika*, no. 2 (45), pp. 50–60. DOI: 10.22213/2618-9763-2021-2-50-60.
- Gippius A.A., 1989. Sistema formalnykh priznakov yazyka drevnerusskoy pismennosti kak predmet lingvisticheskogo izucheniya [The System of Formal Features of the Language of

Ancient Russian Writing As a Subject of Linguistic Study]. *Voprosy yazykoznanija*, no. 2, pp. 93–110.

- GOST R 7.0.64-2018. *Predstavlenie dat i vremeni. Obshchie trebovaniya* [GOST R 7.0.64-2018. Representation of Dates and Times. General Requirements]. Moscow, Standartinform Publ. 40 p. URL: <http://docs.cntd.ru/document/1200159341> (accessed 1 June 2021).
- Mikheev S.M., 2019. Minei dvuh Domok: eshhe raz o pistsah sluzhebnykh minej iz novgorodskogo Lazareva monastyrja [The Menia of Two Domkas: Once More on the Scribes of the Menaia from the Novgorod Convent of St. Lazarus]. *Slověne*, vol. 8, no. 2, pp. 7–56. DOI: 10.31168/2305-6754.2019.8.2.1.
- Molkov G.A., 2020. *Formirovanie orfograficheskikh sistem v drevnerusskoy pismennosti XI – nachala XIII veka: dis. ... d-ra filol. nauk* [Formation of Spelling Systems in Old Russian Writing of the 11th – Early 13th Centuries. Dr. philol. sci. diss.]. Saint Petersburg. 494 p. URL: <https://iling.spb.ru/diss/theses/molkov/thesis.pdf> (accessed 1 June 2021).
- Baranov V.A., Zuga O.V., 2021. Correlation and Cluster Analysis of Fragments of the Earliest Slavonic Gospels. Besters-Dilger J., Rabus A., eds. *El'Manuscript 2021. 8th International Conference on Textual Heritage and Information Technologies (Freiburg im Breisgau, Germany, 12.04.2021–15.04.2021). Booklet of Abstracts*. Freiburg im Breisgau, Albert-Ludwigs-Universität Freiburg, pp. 11–12.
- Höeg C., Zuntz G., eds., 1939. *Monumenta Musicae Byzantinae. Lectionaria. Vol. 1. Prophetologium. Pars. 1. Lectiones Nativitatis et Epiphaniae*. Copenhagen, Haunia, Levin & Munksgaard. 612 p.
- Engberg G., ed., 1980–1981. *Monumenta Musicae Byzantinae. Lectionaria Vol. 1. Prophetologium. Pars. 2. Lectiones anni Immobilis*. Copenhagen, Haunia, Levin & Munksgaard. 314 p.
- Zholobov O.F., Baranov V.A., Novak M.O., 2020. Anonymous vs. Attributed: Cluster Analysis of Tolstovskii Sbornik Texts and Its Interpretation in Terms of Cultural Heritage. *Journal of Siberian Federal University. Humanities & Social Sciences*, 13 (12), pp. 2056–2081. DOI: 10.17516/1997-1370-0704.
- SOURCES**
- Zakhar'inskiy parimeynik 1271 g. [Zacharia's Parimejnik, 1271]. *RNB* [National Library of Russia], Q.p.I.13. 264 l.

- Lazarevskiy (Skovorodskiy) parimeynik, XII v. [Lazar's Parimejnik, 12th Century]. *RGADA* [Russian State Archive of Ancient Acts], f. 381, inv. 1, tip. 50. 126l.
- Troitskiy parimeynik, XIV v. [Troitsa Parimejnik, 14th Century]. *RGB* [Russian State Library], f. 304/I, no. 4. 142l.
- Fedorovskiy II parimeynik, XIII v. [Fyodor II Parimejnik, 13th Century]. *RGADA* [Russian State Archive of Ancient Acts], f. 381, inv. 1, tip. 60. 107l.

CATALOGUES AND DESCRIPTIONS OF HANDWRITTEN COLLECTIONS

- Opisanie slavjanskih rukopisej biblioteki Svjato-Troickoj Sergievoj Lavry. Ch. 1. Svjashhennoe pisanie. № 4 (2005)* [Description of Slavic Manuscripts of the Library of the Holy Trinity Sergius Lavra. Part 1. Holy Scripture. No. 4 (2005)]. Moscow, s. n., 1878. 830 p. URL: <https://www.prlib.ru/item/434785> (accessed 1 June 2021).
- Predvaritel'nyj spisok slavjano-russkih rukopisej XI–XIV vv., khranyashchikhsya v SSSR (dlja «Svodnogo kataloga rukopisej, khranyashchikhsya v SSSR, do konca XIV v. vkljuchitel'no») [Preliminary List of Slavic-Russian Manuscripts of the 11th – 14th Centuries Stored in the USSR (For the

“Consolidated Catalog of Manuscripts Stored in the USSR Until the End of the 14th Century Inclusive”)]. *Arheograficheskiy ezhegodnik za 1965 g.* [Archaeographic Yearbook for 1965]. Moscow, s. n., 1966, pp. 177-273.

- Knjazevskaja O.A., Koval' N.S., Kosheleva O.E., Moshkova L.V. *Katalog slavjano-russkih rukopisnyh knig XI–XIV vv., khranyashchikhsya v TsGADA SSSR Ch. 1* [Catalog of Slavic-Russian Handwritten Books of the 11th – 14th Centuries Stored in the Central State Archive of Ancient Acts of the USSR. Part 1]. Moscow, s. n., 1988. 123 p.
- Svodnyj katalog slavjano-russkih rukopisnyh knig, khranyashchikhsya v SSSR. XI–XIII vv.* [Consolidated Catalog of Slavic-Russian Handwritten Books Stored in the USSR. The 11th – 13th Centuries]. Moscow, Nauka Publ., 1984. 406 p.
- Knjazevskaja O.A., Kobyak N.A., Lifshits A.L., Tikhomirov N.B., Turilov A.A., Shelamanov N.B., eds. *Svodnyj katalog slavjano-russkih rukopisnyh knig, khranyashchikhsya v Rossii, stranah SNG i Baltii. XIV vek. Вып. 1. Apokalipsis – Letopis' Lavrent'evskaja* [Consolidated Catalog of Slavic-Russian Handwritten Books Stored in Russia, CIS and Baltic Countries. The 14th Century. Iss. 1. Apocalypse – Lavrentiev Chronicle]. Moscow, Indrik Publ., 2002. 768 p.

Information About the Author

Victor A. Baranov, Doctor of Sciences (Philology), Professor, Head of the Department of Linguistics, Kalashnikov Izhevsk State Technical University, Studencheskaya St, 7, 426069 Izhevsk, Russia, victor.a.baranov@gmail.com, <https://orcid.org/0000-0003-1730-6359>

Информация об авторе

Виктор Аркадьевич Баранов, доктор филологических наук, профессор, заведующий кафедрой лингвистики, Ижевский государственный технический университет им. М.Т. Калашникова, ул. Студенческая, 7, 426069 г. Ижевск, Россия, victor.a.baranov@gmail.com, <https://orcid.org/0000-0003-1730-6359>