



DOI: <https://doi.org/10.15688/jvolsu2.2022.6.3>

UDC 81'42:26-24
LBC 81.055.1



Submitted: 18.07.2022
Accepted: 19.09.2022

REPRESENTATION AND ANALYSIS OF CONTENT STRUCTURE ELEMENTS IN SAINT PETERSBURG CORPUS OF HAGIOGRAPHIC TEXTS (SCAT)

Elena A. Rogozina

St. Petersburg State University, Saint Petersburg, Russia

Abstract. SCAT is a digital corpus of Old Russian hagiographic texts of 15th–17th centuries created at the Department of Mathematical Linguistics of Saint Petersburg State University. The project involved the introduction of morphological and syntactic markup as well as markup of content structure elements. The first task was to markup meaningful parts (or chapters). The analysis of the texts, which constitute the corpus, allowed to establish a plot structure common to the texts. Markup of the plot components and creation of tables of contents provided an opportunity to work with certain parts of texts and compare similar chapters in different texts. Another task within the framework of the project implied research and markup of biblical, patristic and liturgical quotes in corpus texts. The study of quote distribution in texts and different chapters of the plot enabled the author to reveal the trends in the use of quotes and define the quotes typical of certain parts of the plot. One more task involved markup of repeating text fragments. Research on the texts of SCAT corpus disclosed numerous text excerpts reproduced by different authors. Such extracts are being marked up and links to the original texts are provided. The markup allows to create a register of such fragments, thus simplifying further research of new texts.

Key words: text corpus, hagiography, plot structure, quote, XML markup.

Citation. Rogozina E.A. Representation and Analysis of Content Structure Elements in Saint Petersburg Corpus of Hagiographic Texts (SCAT). *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], 2022, vol. 21, no. 6, pp. 30-46. (in Russian). DOI: <https://doi.org/10.15688/jvolsu2.2022.6.3>

УДК 81'42:26-24
ББК 81.055.1

Дата поступления статьи: 18.07.2022
Дата принятия статьи: 19.09.2022

ПРЕДСТАВЛЕНИЕ И АНАЛИЗ ЭЛЕМЕНТОВ СТРУКТУРЫ СОДЕРЖАНИЯ В САНКТ-ПЕТЕРБУРГСКОМ КОРПУСЕ АГИОГРАФИЧЕСКИХ ТЕКСТОВ (СКАТ)

Елена Андреевна Рогозина

Санкт-Петербургский государственный университет, г. Санкт-Петербург, Россия

Аннотация. СКАТ – электронный корпус агиографических древнерусских текстов XV–XVII вв., созданный на кафедре математической лингвистики Санкт-Петербургского государственного университета. В текстах корпуса вводится морфосинтаксическая разметка, а также активно ведется работа по разметке содержательных элементов. В статье охарактеризованы принципы разметки смысловых разделов. Представлен результат анализа входящих в корпус текстов, позволивший выявить общую для них сюжетную схему. Показано, что разметка сюжетных элементов и создание оглавлений дают возможность работать с отдельными разделами и сравнивать одинаковые разделы в разных текстах. Осуществлен поиск и разметка библейских, святоотеческих и литургических цитат в текстах корпуса. Изучение распределения цитат по текстам и сопоставление цитат с размеченными разделами позволили установить закономерности в использовании цитат и обнаружить цитаты, характерные для определенных разделов житий. Описаны принципы разметки повторяющихся фрагментов текстов. В житиях корпуса СКАТ обнаружены многочисленные текстовые фрагменты, воспроизводимые разными авторами. Такие фрагменты размечаются с указанием ссылки на первоисточник. Эта разметка позволяет постепенно накопить реестр подобных фрагментов и делает возможным их выделение в других текстах.

Ключевые слова: лингвистический корпус, агиография, сюжетная схема, цитата, XML-разметка.

Цитирование. Рогозина Е. А. Представление и анализ элементов структуры содержания в Санкт-Петербургском корпусе агиографических текстов (СКАТ) // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2022. – Т. 21, № 6. – С. 30–46. – DOI: <https://doi.org/10.15688/jvolsu2.2022.6.3>

Введение

СКАТ – электронный корпус агиографических древнерусских текстов XV–XVII вв., созданный на кафедре математической лингвистики Санкт-Петербургского государственного университета (СПбГУ). Работа над ним началась еще в 70-х гг. XX в., и на данный момент корпус содержит более 50 житий общим объемом свыше 500 тысяч словоупотреблений. Более подробную информацию о корпусе СКАТ можно найти на веб-странице, посвященной проекту (<http://project.phil.pu.ru/skat/>), где также представлены 13 житийных текстов в формате XML и PDF. Кроме того, 22 жития общим объемом порядка 185 000 словоупотреблений представлены на веб-странице текстометрической платформы ТХМ лионской лаборатории IHRIM: <https://txm-cactus.huma-num.fr>.

Помимо публикации текстов житий, в рамках проекта ведется исследование собранных материалов и работа по структурированию текста в формате XML. На сайте проекта представлен электронный словоуказатель, который позволяет осуществлять поиск словоформ по всему корпусу текстов (подробно об этой функции см.: [Азарова, Алексеева, 2008]). Для текстов корпуса также вводится морфологическая разметка с указанием не только частей речи, но и типа склонения, рода, падежа и числа для существительных и прилагательных; времени, спряжения, лица, числа для глагольных форм и т. д. Ведется работа и над синтаксической разметкой [Алексеева, Азарова, 2013]. Кроме того, вводится разметка содержательных элементов. Она осуществляется на основе международных норм оформления электронных изданий текста, в частности Text Encoding Initiative (TEI) (<https://tei-c.org/>).

Материал и методы

Занимаясь выявлением и разметкой содержательной структуры текста, следует учитывать отмеченный Д.С. Лихачевым факт, что

композиция житий подчиняется определенным канонам и литературному этикету [Лихачев, 1961]. Помимо трех основных частей: зачина, рассказа о жизни святого и рассказа о совершенных им чудесах, можно выявить схему с более подробным членением, характерную для большинства житийных текстов.

Это обусловлено тем, что при создании агиографического текста зачастую использовался прием сопоставления описываемого святого с его «агиологическим образцом» (или «агиотипом») [Панченко, 2003]: автор жития выбирает одного из уже известных святых, которому и уподобляет героя своего текста. Это сопоставление может осуществляться либо в форме вербального сравнения, либо путем создания нового жития по подобию уже существующего. Готовый текст используется как своего рода шаблон, в котором имеющиеся позиции (второстепенные персонажи, элементы сюжета и т. п.) замещаются новыми образами и заполняются конкретным содержанием. Как отмечал В.О. Ключевский, «получив... образцы агиобиографии, русские слагатели житий однообразно подражали им и в литературных приемах, и в понимании исторических явлений» [Ключевский, 2003, с. 141] и таким образом «житие превратилось в стройное и сложное архитектурное здание, в однообразные формы которого стремились облекать разнообразные исторические явления» [Ключевский, 2003, с. 299]. В результате применения этого подхода сложилась каноническая схема жития, которая переходила из текста в текст.

Следует отметить, что при использовании приема уподобления агиологическому образцу заимствуется не только композиционная схема – из текста с описанием агиотипа могут копироваться целые фрагменты. Иногда они оформляются как явные цитаты, в других случаях обнаружить их можно только при сопоставлении текстов. В текстах житий встречаются цитаты и из других источников, а также устойчивые формулы, сравнения и литературные приемы. Использование подобных топосов весьма характерно для аги-

ографических текстов [Руди, 2005], и зачастую их употребление связано с композицией: те или иные цитаты и клише регулярно появляются в определенных разделах житий.

Учитывая перечисленные особенности житийных текстов, можно ввести три вида содержательной разметки: разметка композиционных элементов (глав), разметка цитат и разметка повторяющихся фрагментов.

В первую очередь для текстов житий вводится разметка композиционных элементов. Зачастую автор жития сам разбивает текст на главы, выделяет части заголовками или буквицами. Однако деление дается не всегда, и в авторской главе может содержаться несколько различных сюжетов. Чтобы обеспечить удобный поиск по текстам и возможность анализа тех или иных элементов, вводится дополнительное деление на смысловые разделы.

Тексты, входящие в корпус СКАТ, представляют собой жития преподобных, и можно было рассчитывать на наличие общих черт в их композиции, поскольку «тип подвига святого... определяет особенности композиционной структуры и поэтики его жития» [Руди, 2006, с. 431]. Характерное для житий преподобных развитие сюжета описывал В.В. Кусков: «Герой происходил, как правило, от благочестивых родителей и с момента своего рождения строго соблюдал посты, чуждаясь детских игр; быстро овладевал грамотой и предавался чтению божественных книг, уединяясь, размышлял о бренности жизни; отказывался от брака, уходил в пустынные места, становился монахом и основывал там обитель; собирал вокруг себя братию, наставлял ее; преодолевал различные бесовские искушения: злокозненные бесы являлись святому в облике диких зверей, разбойников, блудниц и т. п.; предсказывал день и час своей кончины и благочестиво умирал; после смерти тело его оставалось нетленным, а мощи оказывались чудотворными, даруя исцеление недужным» [Кусков, 1982, с. 7].

Опираясь на понимание этих принципов, мы выявили общую для корпуса схему построения сюжета житий, включающую в себя элементы трех уровней.

На первом уровне выделены наиболее крупные элементы сюжета – блоки, в кото-

рых описаны основные события в жизни святого (рождение, учение, уход в монастырь и т. д.). К этому уровню отнесены также введение и заключение. Получена следующая схема:

ВСТУПЛЕНИЕ
 РОДИТЕЛИ
 РОЖДЕНИЕ И МЛАДЕНЧЕСТВО
 УЧЕНИЕ
 ВОЗДЕРЖАНИЕ
 РЕШЕНИЕ УЙТИ В МОНАСТЫРЬ
 МОНАСТЫРЬ
 ОСНОВАНИЕ МОНАСТЫРЯ
 РАСШИРЕНИЕ МОНАСТЫРЯ
 КОНЧИНА
 ПОХВАЛЬНОЕ СЛОВО
 ЗАКЛЮЧЕНИЕ

Далеко не во всех житиях представлены все перечисленные блоки. Например, если отсутствует информация о жизни святого до пострига, то рассказ о родителях, о младенчестве и учении может опускаться.

Некоторые блоки могут повторяться в тексте несколько раз. Прежде всего это относится к той части жития, где описывается жизнь святого после пострига. Например, если святой после пострига переходил из одного монастыря в другой, то в житии может несколько раз повторяться блок «Монастырь».

Каждый из блоков подразделяется на более мелкие компоненты, которые образуют второй уровень. Приведем в качестве примера компоненты, входящие в блок «Монастырь» (см. таблицу).

Так же, как и блоки, отдельные компоненты в конкретном тексте могут опускаться или повторяться при необходимости.

На третьем уровне выделены «плавающие» модули, у которых нет фиксированного места в сюжетной схеме: они могут появляться внутри любых компонентов или блоков либо между ними. Таких модулей пять: «Наставник», «Монашеский подвиг», «Чудо», «Интермедия (риторика / летопись)» и «Быт».

НАСТАВНИК: рассказ о встрече святого с уже известным святым / наставником / архиереем и получении от него благословения; либо, наоборот, о благословении святым другого святого / монаха.

Сюжетная схема житий

Plot structure of hagiographic texts

Блоки уровня 1	Компоненты уровня 2
ВСТУПЛЕНИЕ РОДИТЕЛИ РОЖДЕНИЕ И МЛАДЕНЧЕСТВО УЧЕНИЕ ВОЗДЕРЖАНИЕ РЕШЕНИЕ УЙТИ В МОНАСТЫРЬ МОНАСТЫРЬ	Приход в монастырь Постриг Добродетельное служение Слава и почести Желание уединения и уход из монастыря
ОСНОВАНИЕ МОНАСТЫРЯ РАСШИРЕНИЕ МОНАСТЫРЯ КОНЧИНА ПОХВАЛЬНОЕ СЛОВО ЗАКЛЮЧЕНИЕ	

МОНАШЕСКИЙ ПОДВИГ: рассказ о деяниях святого (противостояние бесовским козням / вразумление или наставление грешников на путь истинный и т. п.).

ЧУДО: описание как прижизненных (например, пророчества), так и посмертных (например, чудесные исцеления над гробом святого) чудес.

БЫТ: описание повседневной жизни монастыря, работы в поле и т. п.

ИНТЕРМЕДИЯ: этот модуль не всегда связан с жизнью святого. Выделяется два подтипа интермедий: авторские рассуждения о событиях жития (риторика) и описание исторических событий того времени, например военных походов (легопись).

Первые четыре модуля, как правило, встречаются в той части, которая повествует о монашеском служении святого. Последний модуль может встречаться в любой части текста.

XML-разметка такой трехуровневой сюжетной схемы позволяет создать для каждого размеченного жития в корпусе своего рода оглавление, с помощью которого можно, как по веб-ссылке, переходить к нужному разделу текста. Это значительно облегчает поиск. Кроме того, появляется возможность сопоставлять и анализировать аналогичные разделы в разных текстах и находить характерные для них языковые особенности, например от-

слеживать использование тех или иных цитат и клише.

В житийных текстах встречается довольно много цитат из различных источников. Основную часть составляют цитаты из Библии, но есть также цитаты из сочинений Отцов Церкви (например, в размеченных текстах корпуса встретились цитаты из Иоанна Златоуста, Василия Великого) и из литургических текстов [Азарова и др., 2019].

Цитаты различаются по размеру, по характеру и по тому, как они вводятся в текст. Все эти особенности учитываются в XML-разметке.

Нами выделены три способа представления цитат в тексте. Во-первых, цитата может вводиться выражением, в котором содержится указание на источник цитаты. Например, *рече господь* указывает на цитату из Евангелия, а *рече пророк Давид* – на цитату из Псалтири. Во-вторых, цитата может быть введена общим выражением, например *якоже пишется* или *сказано бо есть*. В третьем случае цитата формально никак не выделяется в тексте, а органично вписывается в повествование. Для всех этих случаев предусмотрены разные варианты разметки.

В ходе работы мы выделили три типа цитат: точная цитата, видоизмененная цитата и аллюзия.

Точной мы считаем цитату, в которой текст используется в том же виде, что и в пер-

воисточнике. К этому типу также относятся случаи цитирования с небольшими изменениями. Например, когда изменяется время глагола или падеж существительного, чтобы адаптировать цитату и вписать ее в текст жития. Так, в строке *и присно с теми съи их же нога ста на правоте* точно приводится цитата из Псалтири *нога ста на правоте*.

В видоизмененной цитате текст первоисточника может воспроизводиться в сокращенном виде, элементы цитируемого текста могут меняться местами, а отдельные слова могут заменяться на синонимы. Например, во фразе из жития *си суть пощение молитва от чиста сердца... и любовь по Бозе нелицемерна* цитата из 2-го послания к Коринфянам *в любви нелицемерне* адаптирована к контексту.

Третий тип цитат – аллюзия. В этом случае конкретный фрагмент текста первоисточника не цитируется, но дается отсылка к какому-либо его эпизоду. Например, во фразе *но да не и мы впадем в осуждение раба оного, погребшаго в земли талант* содержится отсылка к притче из Евангелия от Матфея, поэтому слова *раба оного, погребшаго в земли талант* мы также размечаем как цитату с указанием типа: «аллюзия».

Для обозначения начала и конца цитат используется тег `<q></q>`, у которого есть ряд удобных для разметки атрибутов. Основным из них является атрибут «source». Он используется в разметке всех типов цитат, потому что позволяет указать «адрес» цитируемого фрагмента. Для Библии это название книги, номер главы и стиха. Например, для приведенной выше цитаты *нога ста на правоте* указывается источник `source="BQ:PSA25:12"`. При цитате-аллюзии указывается весь диапазон стихов, в которых содержится притча, например `source="BQ:MAT25:14-30"`.

Результаты и обсуждение

Совмещая разметку элементов содержания и разметку цитат, мы можем отслеживать, как цитаты распределяются по тексту жития, для каких разделов характерно обилие цитат, а в каких цитаты не используются. Можно также выявлять типичные цитаты, которые характерны для определенного раздела и повторяются из текста в текст (часть цитат размечена на основе данных работы М.К. Кузьминой [2017]).

На данный момент в корпусе СКАТ сюжетная схема и цитаты размечены в четырех текстах: житиях Димитрия Прилуцкого, Григория Пельшемского, Дионисия Глушицкого и Корнилия Комельского. Эта разметка позволяет проводить сравнения и делать определенные выводы.

Прежде всего следует отметить, что тексты значительно различаются по объему и по количеству цитат. В Житии Димитрия Прилуцкого (ДП) зафиксировано около 5 000 словоупотреблений, в Житии Григория Пельшемского (ГП) – около 7 100 словоупотреблений, в Житии Дионисия Глушицкого (ДГ) – около 10 800 словоупотреблений, а в Житии Корнилия Комельского (КК) – около 13 200 словоупотреблений. Однако, как показывают результаты разметки, количество цитат в тексте не зависит от его объема (см. рис. 1).

Диаграмма на рисунке 1 показывает количество цитат в каждом из размеченных текстов. Для каждого жития первый столбец диаграммы демонстрирует общее число цитат в тексте, второй столбец – число цитат из Ветхого Завета (ВЗ), третий столбец – число цитат из Нового Завета (НЗ), а четвертый – число цитат из других источников.

Больше всего цитат встретилось в тексте Жития Григория Пельшемского, которое является далеко не самым большим по объему. В самом большом по количеству словоупотреблений Житии Корнилия Комельского число цитат сопоставимо с числом цитат в самом маленьком по объему Житии Димитрия Прилуцкого.

Совмещение разметки цитат с разметкой сюжетной схемы позволяет сделать некоторые выводы о распределении цитат по смысловым разделам. Использование цитат наиболее характерно для содержательных разделов «Введение» и «Интермедия (риторика)» (рассуждения автора текста), а также для разделов «Добродетельное служение» и «Монашеский подвиг» (в наставлениях преподобного к братии или мирянам часто содержатся цитаты). Реже всего цитаты встречаются в разделах «Чудо» – в них обычно только описывается последовательность событий.

Для дальнейшего анализа текстов мы используем программное обеспечение с открытым кодом ТХМ, разработанное лабораторией IHRIM (Institut d'Histoire des Représentations et

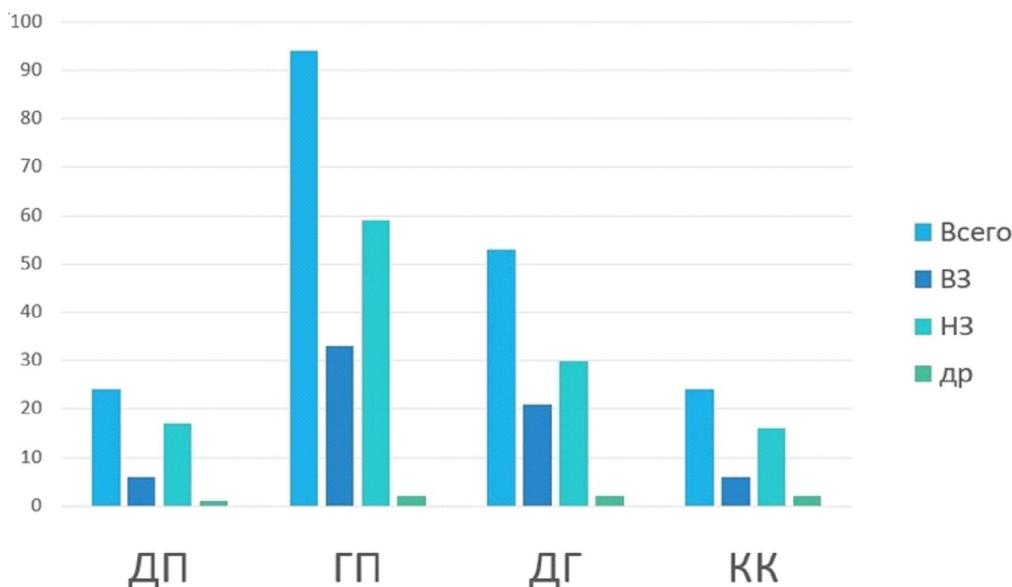


Рис. 1. Количество цитат в текстах житий

Fig. 1. Number of quotes in hagiographic texts

des Idées dans les Modernités) в Лионе. Эта текстометрическая платформа позволяет исследовать размеченные файлы с использованием различных инструментов для выявления закономерностей использования тех или иных элементов структуры содержания. Ознакомиться с ее возможностями можно на веб-портале: <https://txm-sactus.huma-num.fr>. Можно также установить полную версию платформы ТХМ на компьютер и получить доступ ко всем функциям.

Прежде всего ТХМ позволяет представить текст жития с размеченными цитатами и разделами в удобном для пользователя виде; цитаты в тексте подсвечиваются (см. рис. 2).

Далее можно составить конкорданс цитат с помощью запроса на языке CQL (см. рис. 3). Слева указан адрес цитаты в тексте. Если выбрать цитату в конкордансе щелчком мыши, то отобразится визуализация фрагмента жития, в котором выделены все слова цитат.

Помимо собственно представления и поиска платформа ТХМ предлагает и другие инструменты анализа, такие как прогрессия, специфичность и анализ соответствий.

Прогрессия показывает увеличение числа цитат по мере чтения текста (см. рис. 4).

На графике слева показана прогрессия по первому слову в цитатах (то есть по числу цитат). По оси *X* указано количество словоупотреблений. В данном случае оно начинается с цифры 84 000, потому что платформа ТМХ анализирует

весь корпус, а Житие Дионисия Глушицкого, для которого построен этот график, не является первым в корпусе. Показатель $T = 10\ 827$ указывает количество словоупотреблений в данном тексте. По оси *Y* указано количество цитат.

Программа проверяет все слова по направлению чтения текста: от первого слова до последнего. Когда встречается слово, которое является первым словом цитаты, график поднимается на одно деление вверх. Таким образом, мы двигаемся по тексту и отслеживаем, с какой «скоростью» увеличивается число цитат. В тех частях текста, где цитаты следуют одна за другой, график резко идет вверх. Там, где цитат нет, график идет горизонтально.

Вертикальные линии на графике показывают деление текста на основные содержательные разделы – блоки. Сверху рядом с линиями указаны номера этих блоков. Так можно проследить, в каких разделах цитат много, а в каких – мало или нет совсем. Первый раздел Жития Дионисия Глушицкого – это «Введение», где автор обосновывает необходимость написания жития. По графику видно, что в этом разделе довольно много цитат: график резко поднимается вверх. Второй раздел – «Основание монастыря» (в этом житии отсутствует информация о мирской жизни святого), в нем тоже много цитат. В третьем разделе график идет строго горизонтально, то есть в этом разделе (описывающем основание второго монастыря) не встретилось ни одной цитаты.

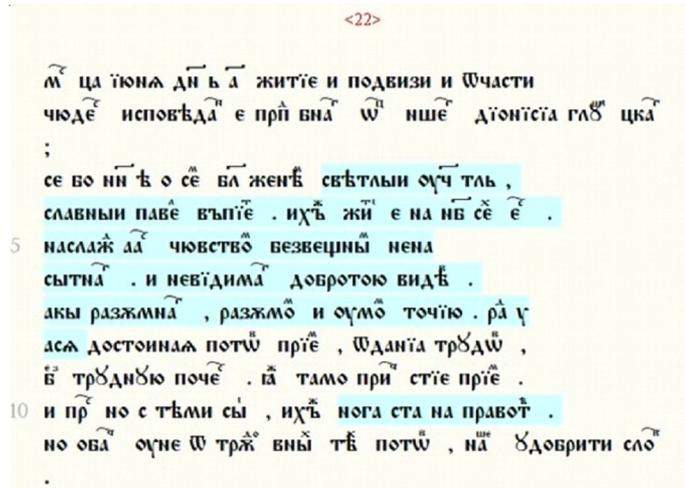


Рис. 2. Визуализация разметки цитат в тексте
 Fig. 2. Visualization of quote mark-up in texts



Рис. 3. Конкорданс
 Fig. 3. Concordance

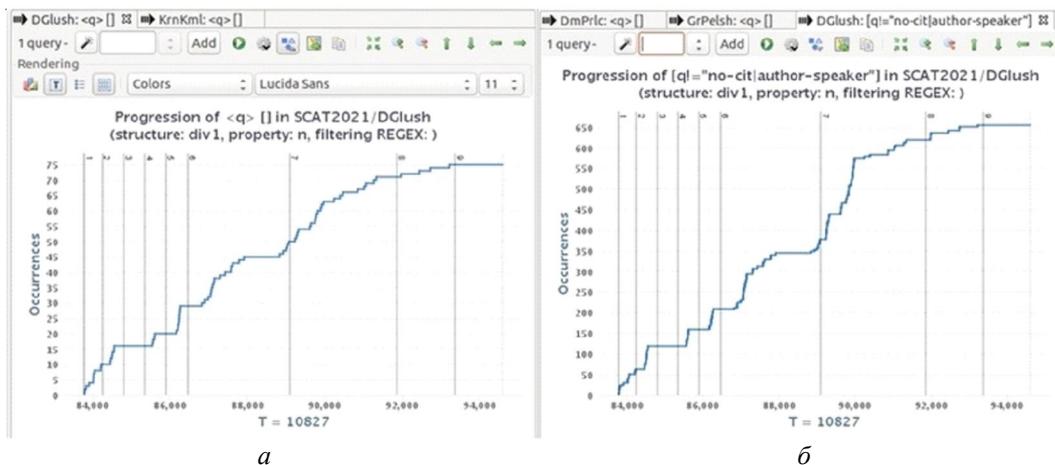


Рис. 4. Прогрессия (для Жития Дионисия Глушицкого):
 а – по первому слову в цитатах; б – по количеству слов в цитатах
 Fig. 4. Progression (for Life of Dionisy Glushitskiy):
 а – by the first word in a quote; б – by the number of words in a quote

На графике справа также представлена прогрессия, но в данном случае она показывает не количество цитат, а количество слов в цитатах. Поэтому показатели на шкале *Y* больше – до 650 слов. В разделах 6 и 7 можно видеть, что график резко поднимается вверх: когда в тексте встречается цитата, график поднимается на столько делений, сколько слов содержится в цитате, то есть резкое движение графика вертикально вверх обозначает длинную цитату.

Таким образом, можно построить отдельные графики прогрессии для каждого текста отдельно и наглядно представить распределение цитат по разным содержательным блокам (рис. 5, 6).

Помимо построения графиков прогрессии платформа ТХМ позволяет создавать частотные словари. Например, можно узнать частоту употребления тех или иных словоформ в цитатах (см. рис. 7).

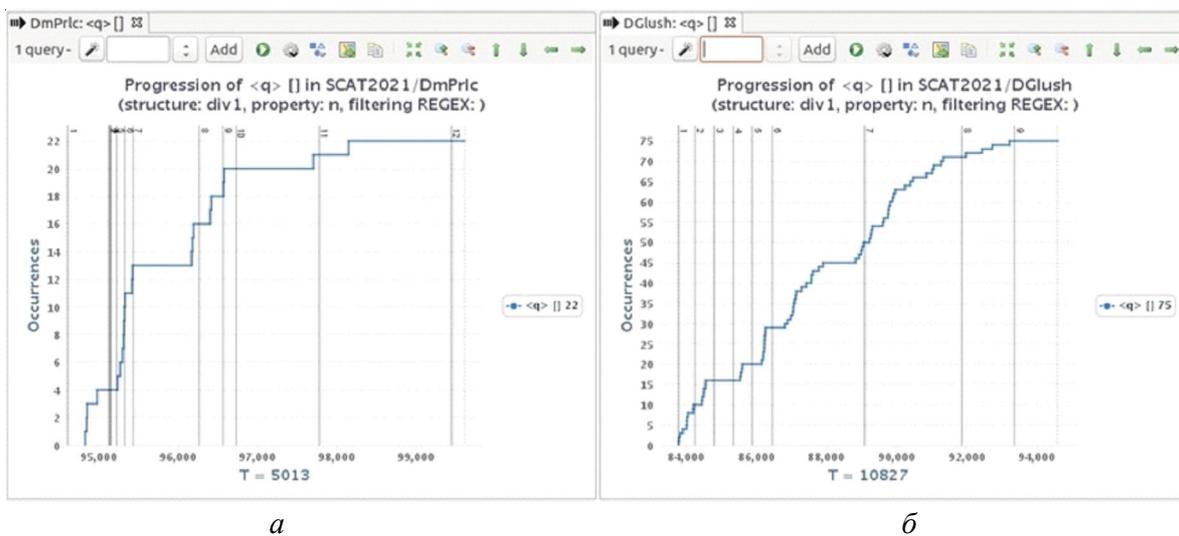


Рис. 5. Прогрессия цитат в житиях:
a – Житие Дмитрия Прилуцкого; *б* – Житие Дионисия Глушицкого
 Fig. 5. Progression of quotes in Lives:
a – Life of Dimitriy Prilutskiy; *b* – Life of Dionisy Glushitskiy

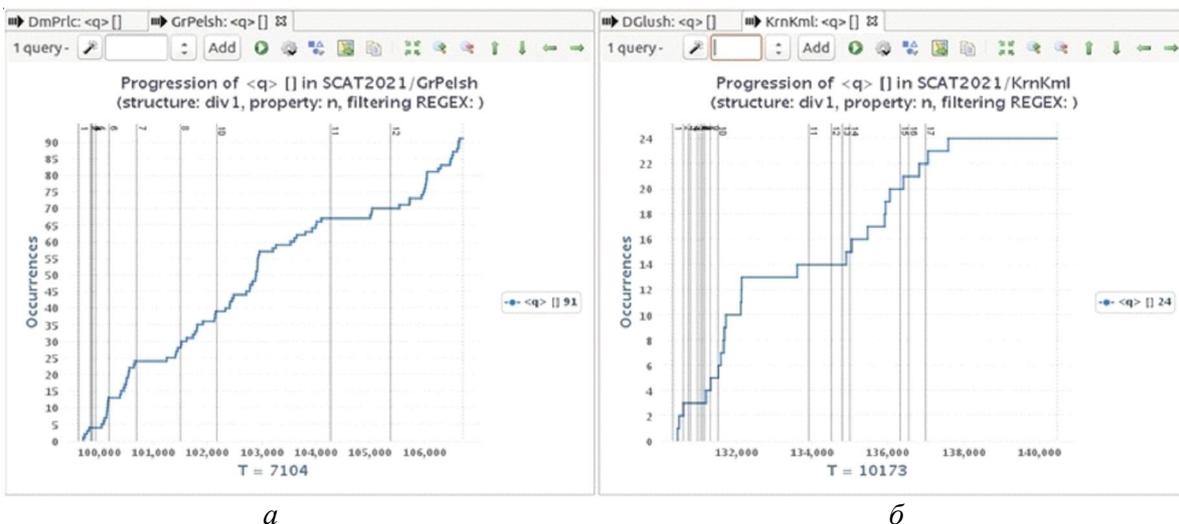


Рис. 6. Прогрессия цитат в житиях:
a – Житие Григория Пельшемского; *б* – Житие Корнилия Комельского
 Fig. 6. Progression of quotes in Lives:
a – Life of Grigoriy Pelshemskiy; *b* – Life of Korniliy Komelskiy

На рисунке 7 в виде таблицы представлены встречающиеся в цитатах словоформы в порядке убывания их частоты (служебные слова исключены из списка). Во втором столбце указывается частота вхождения словоформы во всех размеченных текстах, а в следующих столбцах указаны частоты употребления словоформы в каждом тексте отдельно. Так, можно увидеть, что оказавшееся в начале частотного словаря слово *мене* особенно характерно

для цитат в Житии Дионисия Глушицкого: оно встретилось 15 раз. В остальных текстах оно встречается гораздо реже: в житиях Григория Пельшемского и Корнилия Комельского – по одному разу, а в Житии Димитрия Прилуцкого – ни разу.

Основываясь на данных частотных словарей, платформа ТХМ оценивает специфичность употребления тех или иных словоформ (рис. 8).

word	Frequency T=33117	DGlush t=10827	DmPrlc t=5013	GrPelsh t=7104	KrnKml t=10173
МЕНЕ	17	15	0	1	1
ЕСТЬ	15	6	1	7	1
ИЖЕ	14	5	3	5	1
ЕГО	12	3	0	8	1
МА	9	4	0	5	0
БЪ	7	6	1	0	0
СВОЕГО	7	3	1	2	1
ТИ	7	1	2	4	0

Рис. 7. Частотный словарь словоформ для четырех житийных текстов

Fig. 7. Frequency dictionary of word forms in four hagiographic texts



Рис. 8. Специфичность

Fig. 8. Specificity

Диаграммы на рисунке 8 показывают специфичность первых шести словоформ из частотного словаря в размеченных текстах. При этом отдельно оценивается употребление словоформ в цитатах и в остальном тексте жития. Для этого рассчитывается индекс, который показывает, насколько вероятно столь частое (или редкое) употребление слова в выбранном подкорпусе относительно всего корпуса в целом. Например, индекс специфичности +4 означает, что вероятность настолько частого появления данного слова в выбранном подкорпусе равна $1/10^4$. Значения в диапазоне от -2 до +2 (то есть $1/100$) являются статистически незначимыми. Такие значения могут получиться в результате случайного распределения слов.

На рисунке 8 в верхней части приведена таблица, в которой перечислены словоформы из частотного словаря, указаны частота их употребления и индекс специфичности. Во втором столбце указано, сколько раз эти словоформы встречаются во всех четырех житиях. Далее идут значения для каждого текста отдельно. В третьем столбце указана частота употребления слова в цитатах в Житии Дионисия Глушицкого. В четвертом столбце указан индекс специфичности для такой частоты употребления. В пятом столбце показана частота употребления этого же слова в тексте Жития Дионисия Глушицкого без учета цитат, а в шестом столбце – индекс специфичности для этого числа словоупотреблений в тексте. Аналогичные показатели указаны для всех житий.

В нижней части изображения индексы специфичности представлены в виде столбчатых диаграмм, которые показывают, у каких текстов есть уникальные, присущие только им характеристики. Например, первый столбик в первой диаграмме показывает, что частое употребление слова *мене* является уникальной характеристикой цитат в Житии Дионисия Глушицкого. Индекс специфичности этого слова выше 17, то есть вероятность того, что в подкорпусе такого объема слово *мене* может встретиться 15 раз, крайне мала и составляет $1/10^{17}$. В остальном тексте жития (вне цитат) частота употребления этого слова оказывается в рамках

статистически незначимого диапазона (вторая диаграмма), то есть такое частое использование этой словоформы является уникальной характеристикой только для цитат, содержащихся в этом житийном тексте. Другая ситуация с той же словоформой наблюдается в Житии Корнилия Комельского (две последние диаграммы). В цитатах это слово употребляется в рамках статистической нормы, но для остального текста жития это слово крайне нехарактерно – его показатель специфичности составляет -4. Отрицательное значение показывает, что слово употребляется в тексте гораздо реже, чем могло бы при статистически неспецифичном распределении.

На основе оценки специфичности отдельных слов можно провести анализ соответствий. Это многомерный анализ, который позволяет оценить употребление всех слов во всех разделах текста и определить, насколько тексты близки друг к другу или, наоборот, далеки друг от друга по лексическому составу. Приведем в качестве примера график анализа соответствий для двух житий: Жития Дионисия Глушицкого и Жития Григория Пельшемского (см. рис. 9). В данном случае также отдельно рассматривался лексический состав цитат (DGlushCit и GrPelshCit) и лексический состав остального текста (DGlushNoCit и GrPelshNoCit).

Ожидаемо, тексты двух житий оказались по разные стороны оси 2: лексический состав текстов не совпадает. Однако можно видеть, что лексический состав цитат в обоих житиях значительно отличается от лексического состава остального текста (разделы оказались по разные стороны оси 1). Это различие особенно ярко выражено в Житии Дионисия Глушицкого. Таким образом, результаты многомерного анализа подтверждают, что в цитатах сохраняются языковые особенности оригинала, которые могут не совпадать с языковыми характеристиками основного текста жития, потому что «цитате по определению естественно звучать иначе, чем основному тексту» [Запольская, 2003, с. 482].

Платформа ТХМ предлагает еще один вариант представления анализа соответствий, который позволяет увидеть отдельные совпадающие или специфичные слова (см. рис. 10).

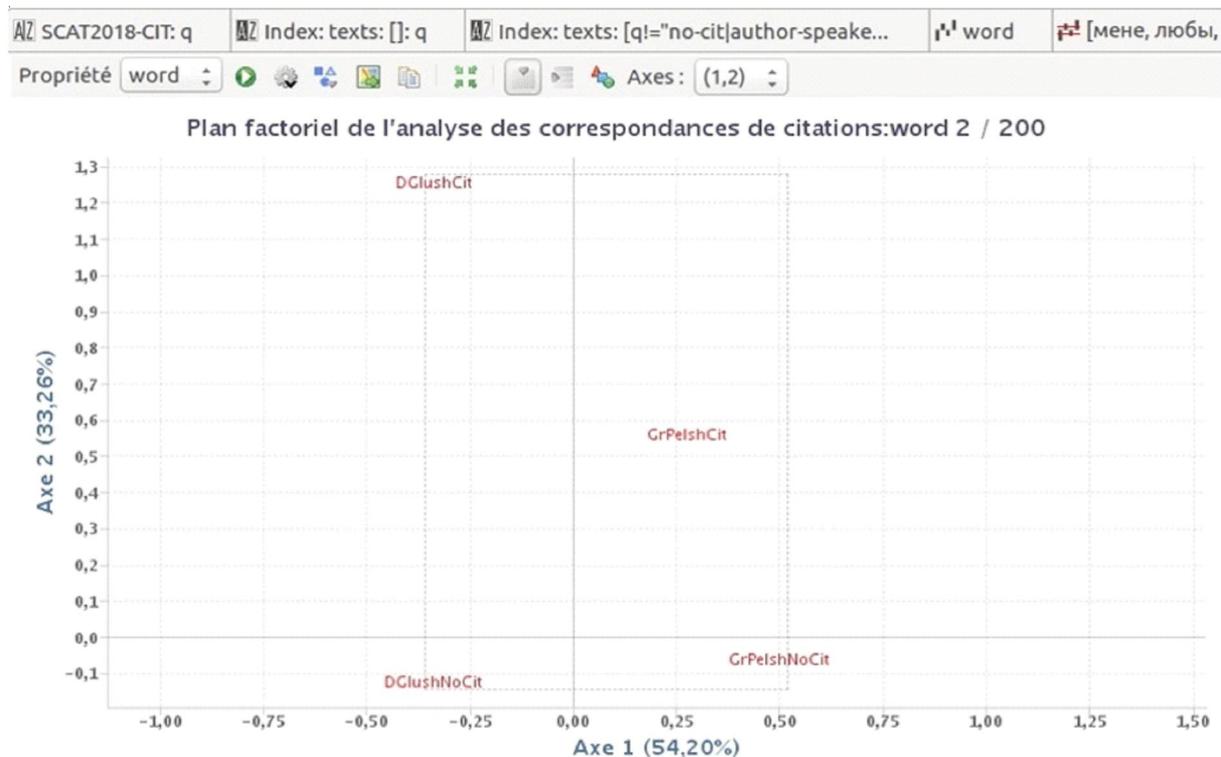


Рис. 9. Анализ соответствий: простое представление
 Fig. 9. Correspondence analysis: simple representation

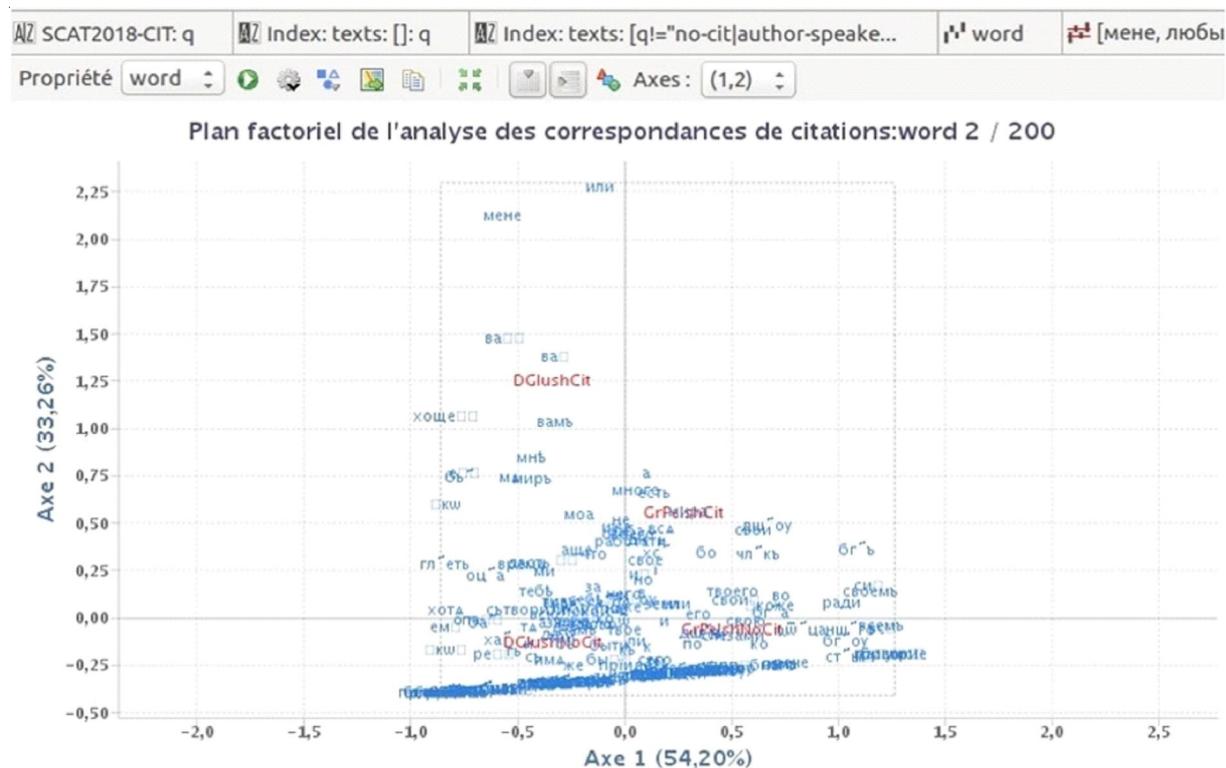


Рис. 10. Анализ соответствий: представление со специфичностью слов
 Fig. 10. Correspondence analysis: representation with word specificity

Слова, имеющие одинаковую специфичность в обоих текстах, накладываются друг на друга, образуя сплошную полосу пересечений в нижней части графика. Чем дальше слово расположено от центра графика, тем более уникальным для данного подкорпуса является его употребление. Например, слово *мене*, которое уже рассматривалось в других примерах, оказалось в верхней части графика, далеко от основной полосы пересечений.

Такие возможности анализа текста, которые дает текстометрическая платформа ТХМ, позволят нам выявлять все новые особенности и закономерности житийных текстов по мере того, как разметка смысловых разделов и цитат будет вводиться в остальных текстах корпуса.

Теперь обратимся к третьему типу разметки элементов содержания – разметке повторяющихся фрагментов для отражения межтекстовых связей в корпусе.

Как уже говорилось, житийные тексты строились по определенному канону и при их написании авторы ориентировались на существующие образцы: в уже накопленном агиографическом материале подбирались жизнеописание святого, близкого по характеру своей деятельности и чертам личности к святому, чье подвижничество предстояло описать, и автор нового жития свободно пользовался текстом предшественника, иногда только заменяя имена и топонимы, иногда адаптируя или сокращая текст или даже заимствуя целые фрагменты текста, «повторяя те же слова и выражения» [Панченко, 2003, с. 507]. Например, в качестве источников для Жития Александра Свирского И. Яхонтов приводит Жития Феодосия Печерского, Сергия Радонежского, Варлаама Хутынского и Кирилла Белозерского [Яхонтов, 1881], уже в наше время Н.В. Пак добавила к этому списку два переводных жития [Пак, 2001].

Можно по-разному относиться к такому использованию чужих сочинений. Например, И. Яхонтов очень низко оценил труд Иродиона, автора Жития Александра Свирского, не признав за ним никакой исторической ценности [Яхонтов, 1881], в то время как Л.А. Дмитриев посмотрел на него с другой точки зрения: «Полнота наших сведений об обстоятельствах написания И<родионом> Жития Алек-

сандра Свирского красноречиво свидетельствует о том, в чем прежде всего агиограф XVI в. видел задачу своего труда: рассказать о святом, имея в виду реальные факты его жизненного пути, во всяком случае хорошо зная эти факты, так, как об этом уже рассказывалось о других святых, предшественниках этого нового святого» [Дмитриев, 1988, с. 441].

Входящие в корпус СКАТ тексты северорусских житий позволяют обнаружить многочисленные текстовые фрагменты, воспроизводимые разными авторами. Поэтому мы решили ввести третий тип структурной разметки: в текстах корпуса отмечены повторяющиеся фрагменты со ссылкой на первоисточник.

Приведем фрагмент, который обнаружился в корпусе СКАТ в житиях Дионисия Глушицкого и Александра Куштского. Он повторяет фрагмент из вступления к Житию Иоанна Лествичника. (Текст Жития Иоанна Лествичника цитируется по публикации Т.Г. Поповой [Попова, 2014].) В результате сопоставления фрагментов можно увидеть, что это один и тот же текст, который заимствуется с некоторыми изменениями (см. рис. 11).

В рамках проекта СКАТ не проводятся текстологические исследования. Решение о том, какой текст считать первоисточником того или иного повторяющегося фрагмента, принимается на основе данных из уже опубликованных работ других исследователей. Например, для фрагмента в данном примере мы указываем в качестве первоисточника Житие Иоанна Лествичника, опираясь на работы Т.Г. Поповой [Попова, 2011]. Если для каких-либо фрагментов со временем будет обнаружен более ранний источник, мы сможем скорректировать разметку и ссылки в корпусе. Выбранная схема разметки позволяет оперативно вносить подобные изменения.

При разметке повторяющихся фрагментов мы даем ссылку на первоисточник. Поскольку во многих случаях цитируемый текст не входит в корпус СКАТ и далеко не все цитируемые тексты представлены в электронном виде, было принято решение добавить оригиналы повторяющихся фрагментов в корпус. Чтобы избежать сложностей, связанных с добавлением полных текстов, мы ограничиваемся только цитируемым фрагментом, со-

Житие Иоанна Лествичника http://www.drevnyaya.ru/vyp/2014_2/part_8.pdf	Житие Дионисия Глушицкого	Житие Александра Куштского
<p>есть во и съ нѣна въ томь. о немъже </p> <p>свѣтлѣи оу читель слави. сде нѣкако</p> <p>въпиа ихъ же житие на нѣсѣхъ есть. насыщаа са [ч]ѣвствѣмь. везв[е]ц[ѣ] нѣмь. ненасытн[а]го и не видим[а]го добротою. [зряа како мьсле] наго разоум[а] [разоу] мѣмь тъчью радоу[а] са. достоино потѣ</p>	<p>се бо нѣѣ о се̄ влжѣнѣ̄</p> <p>свѣтлыи оу чѣтлѣ, слѣвныи пѣвѣ̄</p> <p>въпие̄. ихъ̄ жие̄ на нѣсѣ̄ ѣ̄.] наслаждѣ̄ чювство̄</p> <p>везвѣщны̄ нена сытна̄. ӣ невѣдима̄ добротою видѣ̄.] аки разжѣмна̄, разжѣмо̄ ӣ оумо̄ тѣчю. ра̄дѣ са̄ достоина̄ потѣ̄</p>	<p>се̄и оӯво влжѣныи, о̄ немъ̄̄ вгоглань свѣтлыи оу читель слѣвныи пѣвелѣ̄</p> <p>вопие̄тъ, ихъ̄же жи тїе̄ на нѣсѣхъ̄ ѣ̄сть. наслаждѣ̄ дѣ̄са̄ чювства̄</p> <p>везвѣщны̄ ненасытнаго, ӣ невѣдимаго добротою̄ видѣ̄въ̄ аки разжѣ мна̄ разжѣмомь, тѣчю̄ радѣ̄н са̄ достоина̄ по тебѣ̄,</p>

Рис. 11. Пример повторяющихся фрагментов

Fig. 11. Example of repeating fragments

здаем для каждого отдельный XML-файл и обязательно указываем источник текста, по которому цитируется фрагмент. Именно на этот файл и дается ссылка из текста жития.

При визуализации повторяющийся фрагмент в тексте жития выглядит как веб-ссылка с четко обозначенными границами. При нажатии на ссылку можно перейти к цитируемому оригиналу.

В качестве примера дан повторяющийся фрагмент в тексте Жития Дионисия Глушицкого (см. рис. 12).

Исходный фрагмент в Житии Иоанна Лествичника, представленный в отдельном файле, приведен на рисунке 13.

Приведем еще один пример перехода от повторяющегося фрагмента к первоисточнику (см. рис. 14). Слева представлена страница из Жития Антония Свирского, где также в виде веб-ссылки оформлен фрагмент, описывающий чудо об олени. Справа приведен текст цитируемого фрагмента Жития Евстафия Плакиды. (Мы указываем его в качестве первоисточника, ориентируясь на данные работы О.В. Гладковой [Гладкова, 2013].)

Используя такую разметку, мы составляем реестр повторяющихся фрагментов, что

позволит отслеживать возможные закономерности заимствований и на основе имеющихся данных находить повторяющиеся фрагменты в других текстах житий, а со временем автоматизировать их поиск.

Заключение

Таким образом, в Санкт-Петербургском корпусе агиографических текстов (СКАТ) вводится разметка элементов содержания трех типов: разметка сюжетной схемы, разметка цитат и разметка повторяющихся фрагментов. Совмещение трех типов разметки позволяет анализировать особенности построения житийных текстов. В рамках проекта планируется увеличение объема корпуса и введение разметки в большее число текстов, в том числе разметка наиболее знаковых житий, например Жития Сергия Радонежского, которое послужило образцом для многих других житийных текстов. Это позволит нам выявить общие закономерности в употреблении цитат и заимствований, что в дальнейшем облегчит работу по их поиску в новых текстах и поможет автоматизировать процесс разметки.

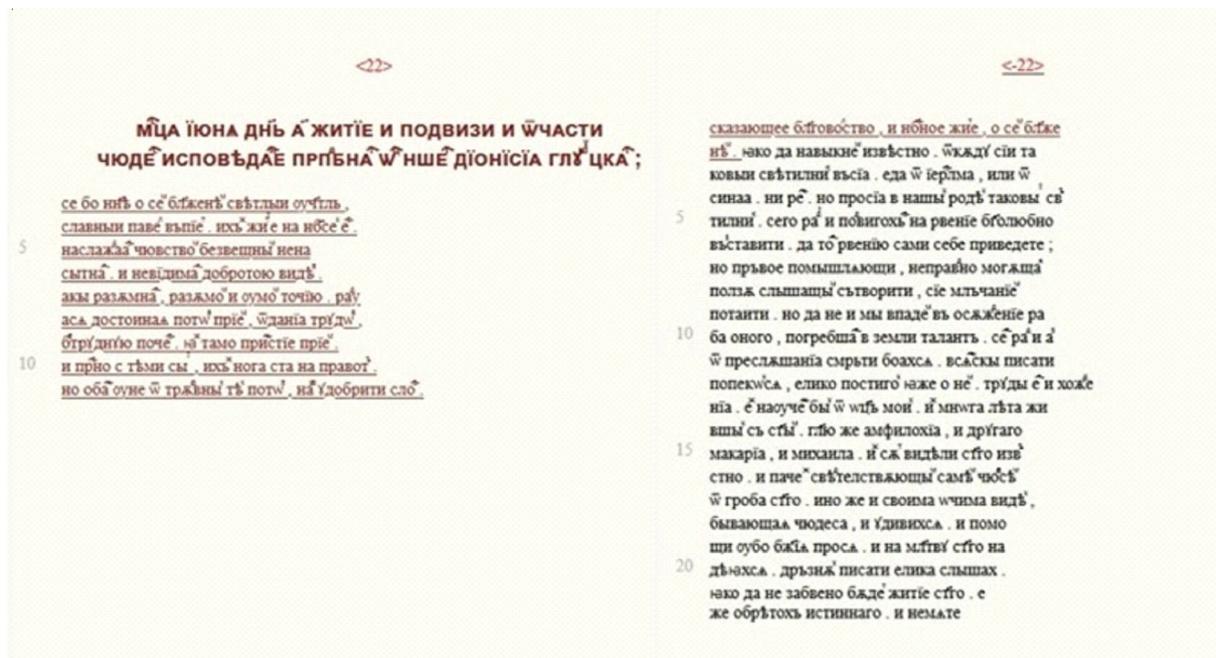


Рис. 12. Визуализация жития Дионисия Глушицкого
 Fig. 12. Visualization of Life of Dionisiy Glushitskiy

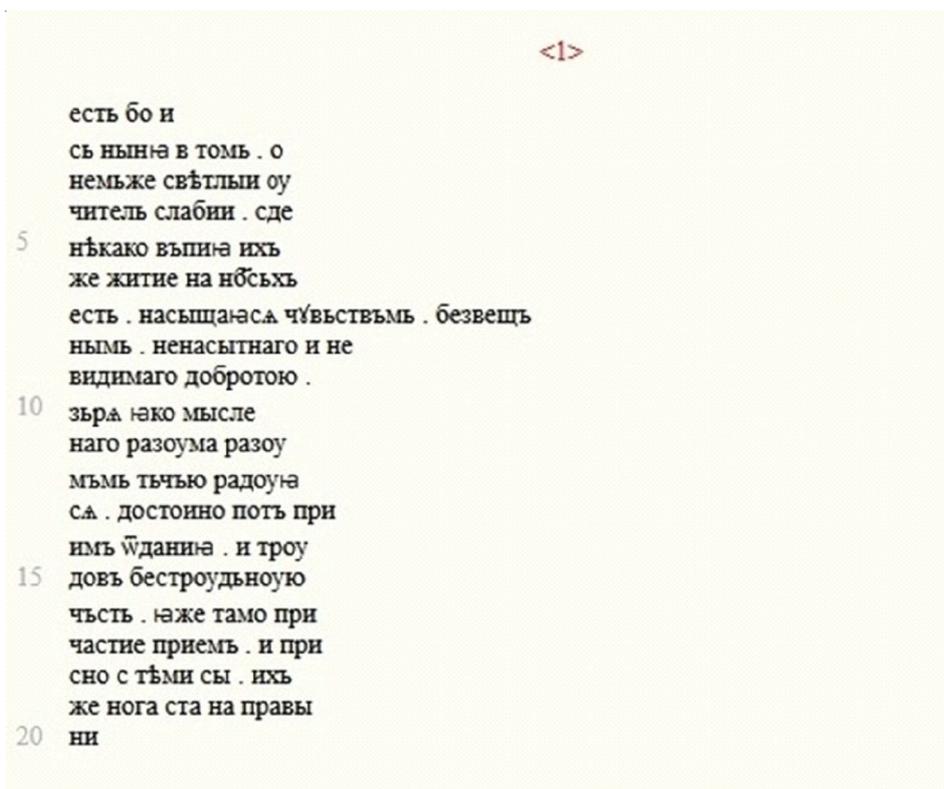


Рис. 13. Исходный фрагмент в житии Иоанна Лествичника
 Fig. 13. Original fragment in Life of John of the Ladder

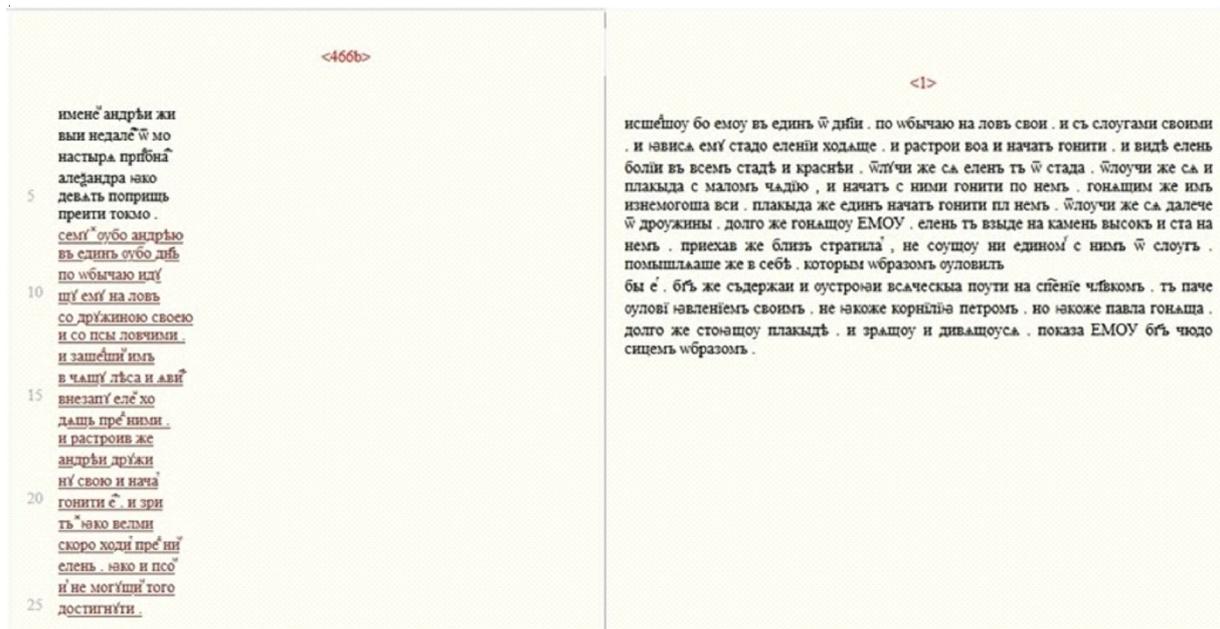


Рис. 14. Фрагмент из жития Евстафия Плакиды в житии Александра Свирского
 Fig. 14. Fragment from Life of Saint Eustathius in the text of Life of Alexander Svirskiy

СПИСОК ЛИТЕРАТУРЫ

- Азарова И. В., Алексеева Е. Л., 2008. Санкт-Петербургский корпус агиографических текстов (СКАТ): формат XML-представления лингвистической информации и организация поиска данных на сайте // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам : Е1'Manuscript-08 : материалы Международ. науч. конф. Казань : Изд-во Каз. ун-та. С. 3–6.
- Азарова И. В., Алексеева Е. Л., Лаврентьев А. М., Рогозина Е. А., Сипунин К. В., 2019. Представление и анализ библейских, святоотеческих и литургических цитат в корпусе СКАТ // Корпусная лингвистика – 2019 : тр. Международ. конф. СПб. : С.-Петербург. гос. ун-т. С. 265–270.
- Алексеева Е. Л., Азарова И. В., 2013. Особенности морфо-синтаксической разметки древнерусских агиографических текстов // Корпусная лингвистика – 2013 : тр. Международ. конф. СПб. : С.-Петербург. гос. ун-т. С. 157–164.
- Гладкова О. В., 2013. Житие Евстафия Плакиды в русской и славянской книжности и литературе IX–XX веков. М. : Индрик. 910 с.
- Дмитриев Л. А., 1988. Иродион // Словарь книжников и книжности Древней Руси. Вып. 2. Вторая пол. XIV–XVI вв. Ч. 1. Л. : Наука, ленингр. отд-ние. С. 440–442.
- Запольская Н. Н., 2003. Библейские цитаты в текстах конфессиональной культуры: семантика, функции, адаптация // Славянский альманах. 2002. М. : Индрик. С. 482–492.
- Ключевский В. О., 2003. Древнерусские жития святых как исторический источник. М. : АСТ : Астрель. 394 с.
- Кузьмина М. К., 2017. Канон преподобнического жития сквозь призму библейских цитат. М. : Водолей. 399 с.
- Кусков В. В., 1982. Литература высоких нравственных идеалов // Древнерусские предания (XI–XVI вв.). М. : Сов. Россия. С. 5–22.
- Лихачев Д. С., 1961. Литературный этикет Древней Руси (к проблеме изучения) // Труды Отдела древнерусской литературы. Л. ; М. : Изд-во Акад. наук СССР. Т. 17. С. 5–16.
- Пак Н. В., 2001. К проблеме источников Жития Александра Свирского // Книжные центры Древней Руси. Севернорусские монастыри. СПб. : Дмитрий Буланин. С. 145–151.
- Панченко О. В., 2003. Поэтика уподоблений (к вопросу о «типологическом» методе в древнерусской агиографии, эпидейктике, гимнографии) // Труды Отдела древнерусской литературы. СПб. : Дмитрий Буланин. Т. 54. С. 491–534.
- Попова Т. Г., 2011. Лествица Иоанна Синайского в славянской книжности. Саарбрюкен : Lap Lambert. 457 с.
- Руди Т. Р., 2006. О композиции и топике житий преподобных // Труды Отдела древнерусской литературы. СПб. : Дмитрий Буланин. Т. 57. С. 431–500.

- Руди Т. Р., 2005. Топика русских житий (вопросы типологии) // Русская агиография. Исследования. Публикации. Polemika. СПб. : Дмитрий Буланин. С. 59–101.
- Яхонтов И., 1881. Жития святых севернорусских подвижников Поморского края как исторический источник. Казань : Тип. ун-та. 377 с.

ИСТОЧНИК

- Попова Т. Г. Житие Иоанна Лествичника (по древнейшей славянской рукописи Лествицы) // Древняя Русь. Вопросы медиевистики. 2014. № 2 (56). С. 83–95.

REFERENCES

- Azarova I.V., Alekseeva E.L., 2008. Sankt-Peterburgskiy korpus agiograficheskikh tekstov (SKAT): format XML-predstavleniya lingvisticheskoy informatsii i organizatsiya poiska dannykh na sayte [Saint Petersburg Corpus of Hagiographic Texts (SKAT): Format of XML Markup for Linguistic Information and Search Function for Web-Site Data]. *Sovremennye informatsionnye tekhnologii i pismennoe nasledie: ot drevnikh tekstov k elektronnyim bibliotekam. El'Manuscript-08: materialy Mezhdunar. nauch. konf.* [Modern Information Technologies and Textual Heritage – from Ancient Texts to Electronic Libraries. El'Manuscript-08. Proceedings of the International Scientific Conference]. Kazan, Izd-vo Kaz. un-ta, pp. 3-6.
- Azarova I.V., Alekseeva E.L., Lavrentyev A.M., Rogozina E.A., Sipunin K.V., 2019. Predstavlenie i analiz bibleyskikh, svyatootecheskikh i liturgicheskikh tsitat v korpuse SKAT. [Representation and Analysis of Biblical, Patristic and Liturgical Quotes in the SKAT Corpus]. *Korpusnaya lingvistika – 2019: tr. Mezhdunar. konf.* [Corpus Linguistics – 2019. Proceedings of the International Conference]. Saint Petersburg, S.Peterb. gos. un-t, pp. 265-270.
- Alekseeva E.L., Azarova I.V., 2013. Osobennosti morfosintaksicheskoy razmetki drevnerusskikh agiograficheskikh tekstov [Aspects of Morphological and Syntactical Mark-up of Old Russian Hagiographic Texts]. *Korpusnaya lingvistika – 2013: tr. Mezhdunar. konf.* [Corpus Linguistics – 2013. Proceedings of the International Conference]. Saint Petersburg, S.-Peterb. gos. unt, pp. 157-164.
- Gladkova O.V., 2013. *Zhitie Evstafiya Plakidy v russkoy i slavyanskoy knizhnosti i literature IX–XX vekov* [Life of Eustathios Plakidas in Russian and Slavic Books and Literature of 9th – 20th Centuries]. Moscow, Indrik Publ. 910 p.
- Dmitriev L.A., 1988. Irodion [Herodion]. *Slovar knizhnikov i knizhnosti Drevney Rusi. Vyp. 2. Vtoraya pol. XIV–XVI vv. Ch. 1* [Dictionary of Scribes and Booklore of Ancient Rus. Iss. 2. The Second Half of 14th – 16th Centuries. Part 1]. Leningrad, Nauka Publ., leningr. otd-nie, pp. 440-442.
- Zapolskaya N.N., 2003. Bibleyskie tsitaty v tekstakh konfessionalnoy kultury: semantika, funktsii, adaptatsiya [Bible Quotes in Texts of Religious Culture: Semantics, Functions, Adaptation]. *Slavyanskiy almanakh. 2002* [Slavic Almanac. 2002]. Moscow, Indrik Publ., pp. 482-492.
- Klyuchevskiy V.O., 2003. *Drevnerusskie zhitiya svyatykh kak istoricheskiy istochnik* [Ancient Russian Lives of the Saints as a Historical Source]. Moscow, AST Publ., Astrel Publ. 394 p.
- Kuzmina M.K., 2017. *Kanon prepodobnicheskogo zhitiya skvoz prizmu bibleyskikh tsitat* [The Canon of the Monastic Life Through the Prism of Biblical Quotations]. Moscow, Vodoley Publ. 399 p.
- Kuskov V.V., 1982. Literatura vysokikh nrvstvennykh idealov [Literature of High Moral Ideals]. *Drevnerusskie predaniya (XI–XVI vv.)* [Old Russian legends (11th – 16th Centuries)]. Moscow, Sovetskaya Rossiya Publ., pp. 5-22.
- Likhachev D.S., 1961. Literaturnyy etiket Drevney Rusi (k probleme izucheniya) [Literature Etiquette of Ancient Rus (on the Issue of Research)]. *Trudy Otdela drevnerusskoy literatury* [Works of the Department of Old Russian Literature]. Leningrad, Moscow, Izd-vo Akad. nauk SSSR, vol. 17, pp. 5-16.
- Pak N.V., 2001. K probleme istochnikov Zhitiya Aleksandra Svirskogo [On the Issue of Sources for Life of Alexander Svirsky]. *Knizhnye tsentry Drevney Rusi. Severnorusskie monastyri* = [Book Centers of Ancient Rus. Monasteries of Russian North]. Saint Petersburg, Dmitriy Bulanin Publ., pp. 145-151.
- Panchenko O.B., 2003. Poetika upodobleniy (k voprosu o «tipologicheskom» metode v drevnerusskoy agiografii, epideyktike, gimnografii) [The Poetics of Assimilation (On the Question of the “Typological” Method in Old Russian Hagiography, Epideictics, Hymnography)]. *Trudy Otdela drevnerusskoy literatury* [Works of the Department of Old Russian Literature]. Saint Petersburg, Dmitriy Bulanin Publ., vol. 54, pp. 491-534.
- Popova T.G., 2011. *Lestvitsa Ioanna Sinayskogo v slavyanskoy knizhnosti* [Ladder of Paradise of John Sinaites in Slavic Literature]. Saarbrücken, Lap Lambert. 457 p.

Rudi T.R., 2006. O kompozitsii i topike zhitiy prepodobnykh [On the Composition and Topoi of the Lives of the Saints]. *Trudy Otdela drevnerusskoy literatury* [Works of the Department of Old Russian Literature]. Saint Petersburg, Dmitryiy Bulanin Publ., vol. 57, pp. 431-500.

Rudi T.R., 2005. Topika russkikh zhitiy (voprosy tipologii) [Topoi of Russian Lives of Saints (the Issue of Typology)]. *Russkaya agiografiya. Issledovaniya. Publikatsii. Polemika*. [Russian Hagiography. Research. Publications. Discussions]. Saint Petersburg, Dmitryiy Bulanin Publ., pp. 59-101.

Yakhontov I., 1881. *Zhitiya svyatykh severnorusskikh podvizhnikov Pomorskogo kraya kak istoricheskiy istochnik* [Lives of the Saints of Russian North as a Source of Historical Information]. Kazan, Tip. un-ta. 377 p.

SOURCE

Popova T.G. Zhitie Ioanna Lestvichnika (po drevneyshey slavyanskoy rukopisi Lestvitsy) [Life of John of the Ladder (Based on the Oldest Slavic Manuscript of the Ladder of Paradise)]. *Drevnyaya Rus. Voprosy medievistiki* [Ancient Rus. Issues of Medieval Studies], 2014, no. 2 (56), pp. 83-95.

Information About the Author

Elena A. Rogozina, Senior Lecturer, Department of Mathematical Linguistics, St. Petersburg State University, Universitetskaya Emb., 11, 199034 Saint Petersburg, Russia, e.rogozina@spbu.ru, <https://orcid.org/0000-0002-5777-1020>

Информация об авторе

Елена Андреевна Рогозина, старший преподаватель кафедры математической лингвистики, Санкт-Петербургский государственный университет, Университетская наб., 11, 199034 г. Санкт-Петербург, Россия, e.rogozina@spbu.ru, <https://orcid.org/0000-0002-5777-1020>