СОЗДАНИЕ И ИСПОЛЬЗОВАНИЕ ИСТОРИЧЕСКИХ КОРПУСОВ СЛАВЯНСКИХ ПИСЬМЕННЫХ ПАМЯТНИКОВ

Виктор Аркадьевич Баранов, д-р филол. наук, профессор, заведующий кафедрой лингвистики Ижевского государственного технического университета имени М. Т. Калашникова, Россия
victor.a.baranov@gmail.com

Аннотация

Рассмотрены требования к функциональным особенностям исторических корпусов средневековых текстов, которые должны быть 1) ориентированы на специфику исходных данных и характер решаемых историко-лингвистических, текстологических и лингвотекстологических задач и 2) реализованы с помощью специальных разметок, процедур обработки, параметров запросов и демонстраций выборок. Корпус должен а) иметь метаданные как о текстах, так и о рукописях, лингвистическую, а также аналитическую разметки, б) обеспечивать демонстрацию документов (сканкопий и транскрипции), конкордансов и перечней и сравнение подкорпусов, в) упрощать графико-орфографическую вариативность при поиске и визуализации данных, г) предоставлять инструменты как для обработки и поиска лингвистического материала и последующего его анализа традиционными методами, так и для постановки и решения задач корпусными методами на основе сведений о количестве, распределении, сочетаемости, вариативности лингвистических единиц в больших массивах данных.

Реализация этих требований демонстрируется на примере подкорпуса трех списков летописей (Лаврентьевского, Ипатьевского, Радзивилловского) исторического корпуса «Манускрипт» (manuscripts.ru).

Сопоставление списка наиболее частотных слов подкорпуса летописей с аналогичными перечнями семи разножанровых подкорпусов показало, что отличительными особенностями первого являются уникальность состава, наличие в нем пяти пространственных предлогов и нек. др. Применение статистических мер TF-ICTF' и Log-Likelihood позволило обнаружить наиболее значимые служебные слова – союз **а**, предлоги **оу**, **о**, **съ**, **къ**, **по**, местоимение **тъ**, наречие **тако**. Использование тех же мер для нахождения статистически значимых знаменательных слов среди пятидесяти наиболее частотных дало возможность показать эффективность применения метода (выделены имена собственные и несколько нарицательных существительных, практически отсутствующих в альтернативных подкорпусах), а также определить статистически значимые нарицательные леммы. Сопоставление статистических значений и рангов лексем позволило найти ядро списков и его ближайшую периферию – **понти**, **дроужина**, **къназь**, **бон**, **протибоу**, **братъ**, **наоуати**, **стати**, **оубити**, **лѣто**, **цьркы**, **възати**, **великъ**, **посълати**, дать их тематическую интерпретацию.

Ключевые слова

Исторический славянский корпус, русские летописи, лингвистическая статистика

CREATION AND USE OF HISTORICAL CORPORA OF SLAVONIC MANUSCRIPTS

Victor A. Baranov, Doctor of Philology, Professor, Head of the Linguistics Department,
Izhevsk State Technical University after M. T. Kalashnikov
Studencheskaya Str. 7, Izhevsk, 426069, Russia
victor.a.baranov@gmail.com

Abstract

The requirements for the operational characteristics of the historical corpora of medieval texts that 1) are determined by the initial data specifics and the character of the historical-linguistic, textological and linguo-textological tasks to be solved 2) and should be realized with the help of special tagging, processing procedures, query parameters and retrieval demonstrations are considered. The corpus should a) have metadata both on texts and manuscripts, linguistic and analytical tagging, b) ensure demonstration of documents (scancopy and transcription), concordances, lists and comparison of subcorpora data, c) simplify graphic-orthographic variation during data search and visualization, d) provide tools both for processing and search of linguistic material and its further analysis by the traditional methods and for problem description and solution by the corpus methods on the basis of data on the quantity, distribution, co-occurrence, variation of linguistic units in great data arrays.

The realization of these requirements is demonstrated on the example of the subcorpus of three copies of chronicles (Laurentian, Hypatian, Radzivilovsky) of the historical corpus "Manuscript" (manuscripts.ru).

The comparison of the list of most frequent words of the subcorpus of chronicles with the similar lists of seven subcorpora of different genres showed that the differential characteristics of the first one is the unique character of its composition, availability of five spatial prepositions and some other. The application of statistic measures TF-ICTF' and Log-Likelihood enabled revelation of the most important empty words – conjunction **а**, prepositions **оу**, **о**, **съ**, **къ**, **по**, pronoun **тъ**, adverb **тако**. The application of the measures to search of statistically important content words among fifty the most frequent words gave a possibility of demonstration of the efficiency of application of the method (proper nouns and some common names that are practically absent in the alternative subcorpora were found), and of determination of statistically important common lemmas. The comparison of the statistic values and ranks of the lexical items helped finding the kernel of the lists and its nearest periphery: lemmas **понти**, **дроужина**, **кънязь**, **бои**, **противоу**, **братъ**, **начати**, **стати**, **оубити**, **лѣто**, **цьркъı**, **възяти**, **великъ**, **посълати**, and giving their thematic interpretation.

Keywords

Historical Slavonic corpus, Russian chronicles, linguistic statistics

1. Ситуация и проблемы

В настоящее время текстовые корпуса современных языков стали незаменимым инструментом для решения различных теоретических и прикладных задач, направленных на выявление речевых и языковых закономерностей, на обеспечение работы поисковых, информационных, аналитических систем, систем перевода, распознавания речи, программ на основе искусственного интеллекта и другого рода высокотехнологичных компьютерных продуктов.

Иная ситуация с созданием электронных копий средневековых рукописных документов, а соответственно – с формированием полнотекстовых коллекций и библиотек на их основе. Несмотря на то, что в последние несколько лет скорость и объемы оцифровки (создания графических образов страниц) рукописей постоянно увеличиваются, предоставляется доступ ко все большему количеству интернет-коллекций и библиотек, содержащих сканированные все с более высоким качеством копии страниц кодексов, снабженные информативными и дружественными по интерфейсу описаниями-каталогами, создание собственно машиночитаемых корпусов в связи с большими трудозатратами по переводу текстов в транскрипции находится в неудовлетворительном состоянии.

Значительно лучше обстоит дело со старопечатными книгами, сканированный образ страниц которых позволяет использовать системы OCR для осуществления поиска лингвистического материала[1].

В Рунете собственно корпусом, содержащим размеченные средневековые славянорусские тексты и снабженным развитыми средствами поиска и демонстрации данных, является Национальный корпус русского языка (ruscorpora.ru). Он имеет, помимо современного материала, древнерусский, старорусский, церковнославянский подкорпуса, содержащие древнейшие и средневековые тексты по спискам разного времени, и подкорпус берестяных грамот 1050–1500 гг.

Интернет-ресурсы, созданные только для работы со славянскими средневековыми документами, немногочисленны (см. список основных машиночитаемых коллекций в конце статьи), а их недостаточная представительность (объем и репрезентативность), зависящая в настоящее время от объективных факторов, затрудняющих подготовку текстовых данных и их разметку, значительно уступает современным корпусам. Коллекции демонстрируют различные технологические, функциональные и интерфейсные решения хранения, разметки, обработки, поиска и демонстрации лингвистических данных.

Если перечислить возможности этих различных по целям и направленности проектов, то в целом получится достаточно большой и разнообразный перечень: они позволяют познакомиться с текстами, просмотреть сканкопии соответствующих страниц рукописей, увидеть вариативность списков одного произведения, найти в текстах интересующие лингвистические формы, получить количественные сведения о них и информацию о их контекстном окружении и сочетаемости (конкордансы, списки n-грамм), при наличии глубокой разметки использовать мета-, аналитические, лингвистические параметры для выявления распределения (частотности) лингвистических единиц или их сочетаний в коллекции (корпусе) и др. Иллюстративные и исследовательские возможности таких коллекций неотделимы друг от друга, так как любая параметризованная выборка, включающая лингвистические единицы, удовлетворяющие некоторым поисковым критериям, уже представляет данные для анализа как минимум традиционными историко-лингвистическими методами[2].

---

[1] Например, интернет-ресурс Google Ngrams Viewer (https://books.google.com/ngrams), судя по временной школе, обеспечивает поиск словоформ и их сочетаний в книгах начиная с 1500.

[2] При необходимости исследовать средневековый текст как лингвистический феномен мы, как известно, лишены возможности опираться на личный или социальный речевой опыт и можем делать выводы только на основании имеющихся текстовых фактов в сохранившихся до нашего времени рукописях.

Несомненно, перспективным является использование исторических корпусов в качестве источников количественных и статистических сведений, позволяющих выявлять частотность лингвистических единиц, их сочетаемость с другими единицами и распределение в текстах (в конкретных списках) в сравнении с ожидаемыми средними величинами и частотой других единиц. Такие данные могут как способствовать решению некоторых традиционных лингвистических задач, так и помочь сформулировать нестандартные для исторической лингвистики проблемы, решение которых без привлечения большого массива материала невозможно.

Эти два направления использования корпусов – традиционное, требующее лингвистического анализа контекстов, учета бытования текстов, необходимости критики текста и других исследовательских процедур, и корпусное, основанное на формальной параметризации данных и нацеленное на поиск закономерностей в использовании лингвистических единиц, – не противоречат и тем более не противопоставлены друг другу[3], а дополняют друг друга своими результатами. Более того, историко-лингвистические исследования, объектом которых являются в первую очередь альтернативные формы выражения некоторого значения, используют по сути корпусный метод, так как при анализе исследователь в идеале должен учесть все количество форм, их сочетаемость и распределение в тексте (подкорпусе) [Баранов 2015, с. 41]. Поэтому «ручные» и «машинные» методы исторического языкознания не имеют принципиальных, сущностных, содержательных различий.

Исторический корпус, создание, поддержание и развитие которого является особой технологической и лингвистической задачей, – это одновременно и цель работы лингвистов и программистов (но не самоцель, не конечный самодостаточный результат), и инструмент для поиска закономерностей в реализации речевых форм и структур (но, конечно, отнюдь не единственный аналитический инструмент).

2. Исторический корпус как цель и инструмент

В работе [Баранов 2015] рассмотрены требования к историческим корпусам с точки зрения хранения, разметки поиска, демонстрации и анализа текстовых данных:

– поиск документов не только по метаданным произведений, но и по характеристикам рукописей как их физических носителей,

– создание не только мета- и лингвистической разметки, но и аналитической, дающей возможность анализировать части (фрагменты) произведений и списков,

– обеспечение поиска и визуализации данных с различной степенью точности по отношению к оригиналу,

– визуализация выборок в соответствии с задачами пользователя (контекст, указатели, конкордансы, перечни сочетаний, параллельные корпуса и др.),

– демонстрация не только контекста (в том числе в виде конкорданса), но и адреса лингвистической единицы,

– предоставление возможности найти дефектные части рукописи,

– обеспечение демонстрации сканированных копий страниц,

[3] Ср., например, регулярно возобновляемую полемику относительно методов гуманитарных исследований, в частности в филологии, где традиционные методы (close reading – «медленное, близкое» чтение) непримиримо противопоставляются корпусным, реализующимся на основе материалов текстовых баз данных (distant reading – «быстрое, дистантное» чтение). (Термины *close* / *distant reading* введены Франко Моретти (Franco Moretti), Стэнфордский университет [Аллингтон и др. 2017].)

– наличие специализированных инструментов для корпусных исследований (количественного, статистического, дистрибутивного и позиционного анализа) и др.

Несмотря на неизбежный сегодня относительно небольшой объем исторических корпусов, их возможности уже сегодня заключаются (или должны заключаться) не только и не столько в быстром нахождении данных, сколько в обеспечении поиска различий между подкорпусами, обладающими различными мета- или аналитическими характеристиками, в сопоставлении данных, извлеченных из разновременных списков и/или текстов разных типов (жанров), то есть обеспечивать выявление таких различий, которые являются реализацией речевой (и языковой) вариативности средневековых текстов и их списков, и нахождение закономерностей в распределении вариантов.

Этими целями и задачами и должны определяться особенности модели данных, принципы мета-, аналитической и лингвистической разметки, функциональные возможности инструментов доступа и анализа, параметров поиска и форм демонстрации результатов запросов в исторических корпусах.

3. Исторический корпус «Манускрипт: славянское письменное наследие» как инструмент демонстрации и анализа данных

В настоящее время инструментарий корпуса “Манускрипт” (модули поиска и демонстрации, правила приравнивания вариантов, электронный грамматический словарь и автоматический лемматизатор, процедуры поиска точных и неточных соответствий и мн. др.) предназначен для решения указанных задач. Он позволяет ставить цели, достижение которых возможно в том числе и с помощью корпусных методов.

Модули корпуса обеспечивают поиск лингвистического материала в славянских разножанровых рукописях X–XV вв. и его демонстрацию в виде упорядоченных по различным критериям перечней словоформ и лемм, их контекстов, параллельных фрагментов списков одного произведения, статистической и дистрибутивной оценке текстовых единиц как в одном списке, так и в нескольких и др., то есть в таких формах визуализации, которые позволяют получить данные для традиционного и корпусного анализа текстовых данных.

4. Статистический анализ лексики средневековых текстов: цели и задачи

Количественно-статистические методы давно с успехом используются для анализа современных текстов с целью создания поисковых, аналитических, интеллектуальных систем различной направленности.

Не менее результативным и эффективным, как мы полагаем, может быть использование таких методов при рассмотрении лингвистического материала исторического корпуса.

Продемонстрируем возможности количественных и статистических методик на материале подкорпуса русских летописей, включив в него три древнейших списка – Лаврентьевский (ЛЛ), Ипатьевский (ЛИ) и Радзивилловский (ЛР).

Цель работы – с помощью количественного и статистического анализа сопоставить наиболее частотные слова летописных текстов с соответствующими перечнями текстов других жанров корпуса.

Задачи работы:

– выявить наиболее частотные слова подкорпуса,

– с помощью статистических мер TF-ICTF' и LL определить наиболее значимые лексические единицы,
– сопоставить их ранги и выявить ядро и периферию перечней,
– дать лингвистическую интерпретацию ключевых слов.

5. Материал, методика, приемы & обсуждения

5.1. Объем данных

Объем подкорпуса летописей – 349 913 словоформ.

Контрастный (альтернативный) корпус включает данные 38-ми списков XI–XIV вв. разножанровых текстов (7 списков майской минеи, 6 – миней на другие месяцы года, 5 – стихираря, 11 – Евангелий, 2 – Апостола, 3 – Паренесиса, 4 – Псалтыри), общий объем корпуса – 1 369 305 словоформ.

5.2. Материал

Исследованию подвергнуты наиболее часто встречающиеся в подкорпусе летописей слова.

Перечень наиболее часто встречающихся слов, сформированный с помощью модуля n-грамм корпуса «Манускрипт», приведен в таблице 1.

Таблица 1. Список 10-ти наиболее часто встречающихся в летописях слов

№	*Слово*[4]	*Абсолютное количество*	*Относительное количество*
1	**н**	30615	0,087
2	**сѧ**	13072	0,037
3	**же**	12702	0,036
4	**въ**	9825	0,028
5	**на**	6147	0,018
6	**не**	5117	0,015
7	**къ**	4781	0,014
8	**съ**	4742	0,014
9	**а**	4366	0,012
10	**о҃т҃ъ**	2598	0,007

Соответствующие списки привлеченных для сравнения подкорпусов даны в приложении 1.

5.3. Первое обсуждение

Сопоставление таблиц дает возможность увидеть, что во всех подкорпусах наиболее частым словом является союз-местоимение **н**. Среди первых 10-ти слов летописей имеются и другие, которые совпадают с наиболее частотными словами других подкорпусов, – **сѧ**, **же**, **въ**, **не**. В то же время уникальными в перечне первых 10-ти слов подкорпуса летописей являются предлоги **къ**, **съ** и союз **а**.

Второе отличие заключается в наличии среди первых 10-ти слов 5-ти пространственных предлогов: **въ**, **на**, **къ**, **съ**, **о҃т҃ъ**. Третье – в отсутствии в списке местоимений (кроме **сѧ**), союза **яко** и форм глагола *быти*, которые входят в перечни всех других подкорпусов. Следует также отметить, что количество совпадений с другими корпусами практически одинаково и не превышает 6–7-ми единиц, в то

[4] Омонимия не снята. Учтены имеющиеся в текстах варианты слов, например: для **н** – **н**, **ı**, **ї**, для **сѧ** – **сѧ**, **сꙗ**, для **же** – **же**, **жє**, **жѥ**, для **въ** – **въ**, **въ**, **во**, **в** и т. п. В связи с возможной различной передачей в транскрипции некоторых компонентов (**сѧ**, **же**, **не**) учитывались оба варианта передачи – слитный и раздельный.

время как между другими подкорпусами оно иное – от 7 (например, между Евангелиями и Апостолами) до 10 (например, между минеей на май и минеями на другие месяцы года).
Уже такое простое сопоставление позволяет обнаружить несколько отличительных черт незнаменательной лексики летописей, часть из которых ожидаема, например, большое количество пространственных предлогов в 10-ти первых позициях, часть – не очевидна, например, отсутствие в списке некоторых наиболее частотных слов других корпусов или примерно равнозначные по количеству совпадения с другими подкорпусами.

5.4. Статистические методы

Для выявления слов, наиболее значимых в перечне первых 10-ти, использованы две статистические меры – TF-ICTF (вариант меры TF-IDF) и Log-Likelihood[5].
Оба метода используются в лингвистической статистике для выявления претендующих на статус ключевых слов документа, а также для оценки близости документа другим документам корпуса.

5.4.1. Мера TF-ICTF

Мера TF-ICTF (term frequency – inverse collection term-frequency) является одним из вариантов широко известной меры TF-IDF (term frequency – inverse document frequency)[6] и предполагает учет как частотности лингвистических единиц в анализируемом документе, так и их частотности в корпусе в целом[7]:

$$TF\text{-}ICTF = \frac{f_d}{F_d} * \log\frac{F_D}{f_D},$$

где f_d – количество анализируемых словоформ/лемм (term) в документе (в нашем случае – в коллекции),
F_d – количество всех словоформ/лемм в документе (коллекции),
F_D – общее количество словоформ/лемм во всех документах корпуса (во всех коллекциях),
f_D – количество анализируемых слов/лемм во всех документах корпуса (во всех коллекциях).

Эксперименты показали, что статистический вес слова в соответствии с этой мерой существенно зависит от частоты его использования в анализируемом документе: чем чаще употребление лексической единицы, тем выше ее значение даже в том случае, когда слово часто встречается в корпусе в целом.
Для уменьшения влияния на вес слова частоты его встречаемости в документе и повышения роли частоты слова в корпусе используются различные модификации меры.
Эксперименты показали, что одним из приемлемых вариантов является мера TF-ICTF' с дополнительными коэффициентами:

$$TF\text{-}ICTF' = (0{,}5 + 0{,}5\frac{f_d}{F_d}) * \log\frac{F_D - F_d}{f_D - f_d},$$

---

[5] Статистические расчеты сделаны в программе MS Excel.
[6] Мера TF-IDF используется для выявления специфики документа и поиска в нем уникальных слов на основе сравнения имеющихся в документе слов с их наличием/отсутствием в других документах корпуса. О мере см., например, [Sparck 1972; Salton and Yang 1973; Robertson 2004; Roelleke 2013].
[7] О мере TF-ICTF см., например, [Kwok 1995; Roelleke and Wang 2006; Wu et al 2008, p. 17–18].

где $F_D - F_d$ – объем всех коллекций без объема коллекции, в которую входит лемма, для которой вычисляется вес,

$f_D - f_d$ – количество леммы во всех коллекциях, кроме количества в той коллекции, в которую входит анализируемая лемма.

Значения меры TF-ICTF' для 10-ти первых слов подкорпуса летописей приведены в таблице 2.

Таблица 2. Статистический вес 10-ти наиболее часто встречающихся в летописях слов (мера TF-ICTF')

№	*Слово*	*Абсолютное количество*	*TF-ICTF'*
1	н	30615	0,6451
2	сѧ	13072	0,8151
3	же	12702	0,7868
4	въ	9825	0,8243
5	на	6147	1,0124
6	не	5117	0,8345
7	къ	4781	1,1284
8	съ	4742	1,1505
9	а	4366	1,4003
10	ѿ	2598	0,9577

5.4.2. Мера Log-Likelihood

Не менее часто используемой статистической мерой для нахождения слов, претендующих на статус ключевых, и на этом основании – для выявления отличий документа от других документов корпуса является мера Log-Likelihood (коэффициент правдоподобия, показатель сходства)[8]:

$$LL = 2(a \ln\left(\frac{a}{c\frac{a+b}{c+d}}\right) + b \ln\left(\frac{b}{d\frac{a+b}{c+d}}\right)),$$

где a – абсолютное количество анализируемой словоформы/леммы в документе (коллекции),

b – абсолютное количество анализируемой единицы в других документах (коллекциях) корпуса,

c – объем документа (коллекции), в котором анализируется единица,

b – объем других документов (коллекций) корпуса.

Значения меры LL для 10-ти первых слов подкорпуса летописей приведены в таблице 3.

Таблица 3. Статистический вес 10-ти наиболее часто встречающихся в летописях слов (мера LL)

№	*Слово*	*Абсолютное количество*	*LL*
1	н	30615	937,606
2	сѧ	13072	984,381
3	же	12702	79,405
4	въ	9825	6,941
5	на	6147	814,073
6	не	5117	1379,604
7	къ	4781	1534,979

[8] О мере см., например, [Rayson, Garside 2000, p. 3; Ляшевская, Шаров 2009, стр. 8–9].

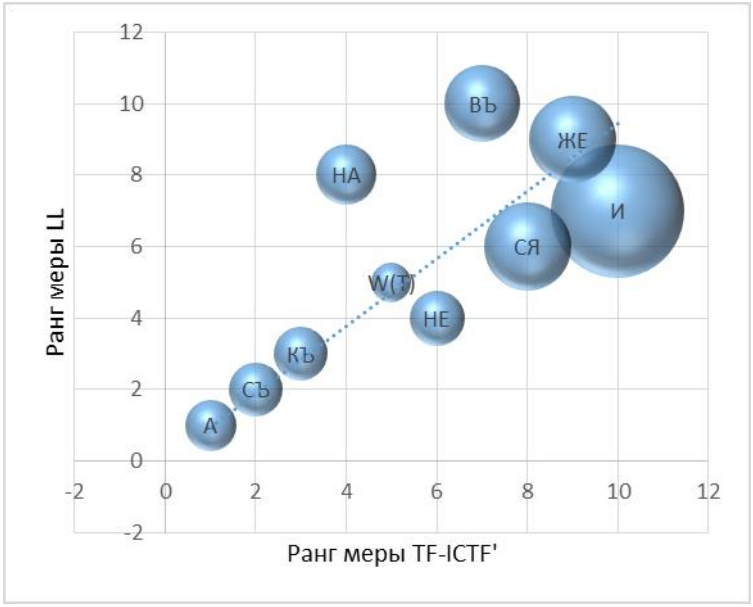
№	Слово	Абсолютное количество	LL
8	сѣ	4742	1855,514
9	а	4366	5754,154
10	ѿ	2598	988,316

5.5. Второе обсуждение

Результаты измерения двумя мерами во многом идентичны: наибольший вес имеют союз а, а также предлоги сѣ и кѣ, наиболее часто используемые слова н, са, же, вѣ имеют наименьший вес.

Представим значимость служебных слов с помощью диаграммы, на которой в качестве осей даны ранги слов, а величина шара соответствует относительному количеству слова.

Диаграмма 1. Ранг меры TF-ICTF' & ранг меры LL & относительное количество 10-ти наиболее частотных слов



Слова а, сѣ и кѣ составляют своеобразное ядро и оцениваются мерами абсолютно одинаково, о чем свидетельствуют их идентичные ранги (1, 2 и 3 соответственно) и их нахождение на линии тренда диаграммы. Оценка слов н, са, с одной стороны, и вѣ, на, с другой, несколько различается: первые имеют более высокий ранг в соответствии с мерой LL (7 и 6 vs 10 и 8 по мере TF-ICTF'), вторые – более высокий ранг по мере TF-ICTF' (7 и 4 vs 10 и 8 по мере LL).

5.6. Статистические значения 25-ти наиболее часто использующихся слов

Рассмотрим, что происходит при увеличении количества анализируемых слов корпуса. Приведем список 25-ти наиболее частотных форм подкорпуса летописей, их относительное количество, статистические значения и ранги в соответствии с ними, см. таблицу 4.

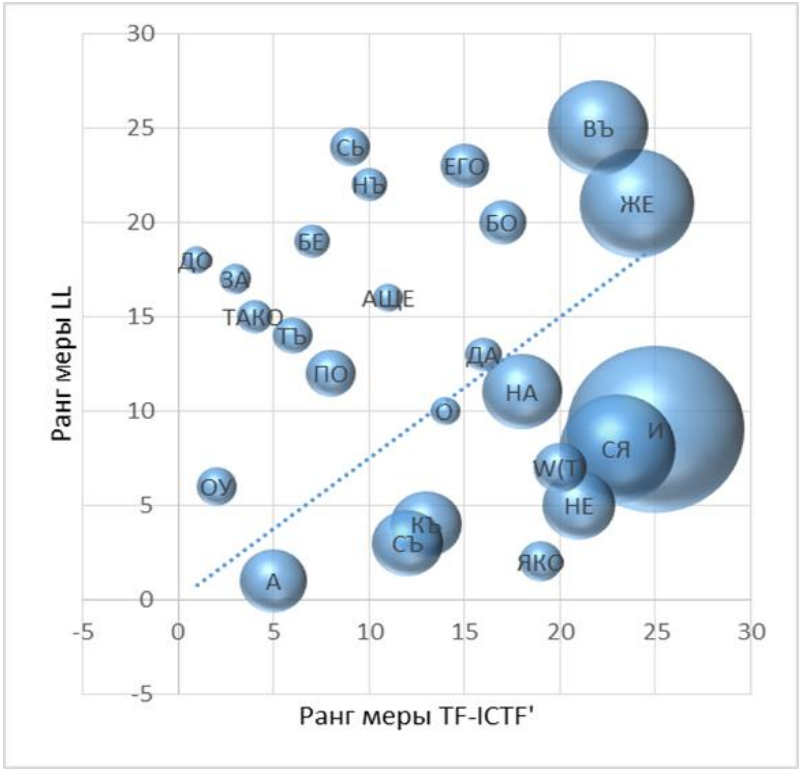
Таблица 4. Статистический вес 25-ти наиболее часто встречающихся в летописях слов (меры TF-ICTF' и LL) и их ранги

№	Слово/форма	Абсолютное количество	Относительное количество	TF-ICTF'	Ранг	LL	Ранг
1	н	30615	0,087	0,64513	25	937,6056	9
2	са	13072	0,037	0,81511	23	984,3811	8

№	*Слово/форма*	*Абсолютное количество*	*Относительное количество*	*TF-ICTF'*	*Ранг*	*LL*	*Ранг*
3	же	12702	0,036	0,78678	24	79,4046	21
4	бъ	9825	0,028	0,82429	22	6,9411	25
5	на	6147	0,018	1,01240	18	814,0733	11
6	не	5117	0,015	0,83447	21	1379,6043	5
7	къ	4781	0,014	1,12841	13	1534,9793	4
8	съ	4742	0,014	1,15049	12	1855,5140	3
9	а	4366	0,012	1,40034	5	5754,1537	1
10	ѿ	2598	0,007	0,95767	20	988,3155	7
11	по	2388	0,007	1,25623	8	639,1238	12
12	его	2147	0,006	1,10468	15	28,3746	23
13	бо	2093	0,006	1,08700	17	97,1155	20
14	яко	1737	0,005	0,95776	19	2093,1471	2
15	се/сь[9]	1519	0,004	1,23314	9	17,7249	24
16	оу	1509	0,004	1,50024	2	1244,3498	6
17	то/тъ	1417	0,004	1,39770	6	527,4923	14
18	да	1224	0,003	1,09969	16	622,7685	13
19	тако/такъ	1147	0,003	1,46063	4	501,2809	15
20	бѣ/бе	1165	0,003	1,36935	7	180,0803	19
21	но/нъ	1133	0,003	1,21442	10	60,0578	22
22	за	921	0,003	1,47668	3	301,8552	17
23	аще	867	0,002	1,19506	11	312,3113	16
24	ѡ/о	844	0,002	1,12211	14	911,4090	10
25	до	745	0,002	1,53650	1	281,4606	18

Данные представим в виде диаграммы 2.

Диаграмма 2. Ранг меры TF-ICTF' & ранг меры LL & относительное количество 25-ти наиболее частотных слов (форм)



[9] На первом месте в списке стоит наиболее частотный вариант формы.

5.7. Третье обсуждение

Ядром перечня являются союз **а** и предлог **оу**: обе меры размещают их в пределах первых 10-ти рангов. Ближайшая периферия ядра (ранги с 11 по 15) – предлоги **о**, **съ**, **къ**, **по**, местоимение **тъ**, наречие **тако**. Видно, что и при увеличении количества анализируемых слов среди наиболее статистически значимых – в первую очередь предлоги.

Одинаково низко меры оценивают предлог **въ** и частицы **же** и **бо**, которые входят в список наиболее частотных слов всех корпусов (см. приложение 1).

В то же время оценка других слов статистическими мерами существенно отличается: наиболее часто встречающиеся в подкорпусе слова мерой LL оцениваются с точки зрения значимости высоко (на диаграмме смещены вправо и вниз), менее часто встречающиеся – как незначимые (на диаграмме смещены влево и вверх). Различия в статистических значениях мер TF-ICTF' и LL демонстрирует также смещенная вправо – в сторону наиболее частых слов (большой размер шаров) – линия тренда.

В целом две использованные статистические меры успешно справляются с поиском в подкорпусах форм, отличающих подкорпус летописей от текстов других жанров, и позволяют выявить в подкорпусе значимые незнаменательные единицы.

5.8. Анализ лемм

Используем методику сопоставления статистических значений лингвистических единиц для исследования знаменательных слов.

Материал

Материалом для анализа явились наиболее часто встречающиеся в подкорпусе летописей слова (леммы) знаменательных частей речи. Количество лемматизированных словоформ в подкорпусе – 263 426[10].

Перечень наиболее частых 50-ти лемм приведен в таблице 5.

Таблица 5. Список 50-ти наиболее часто встречающихся в летописях знаменательных слов, их статистический вес (меры TF-ICTF' и LL) и ранги

№	*Лемма*	*Абсолютное количество*	*Относительное количество*	*TF-ICTF'*	*Ранг*	*LL*	*Ранг*
1	**бытн**	5676	0,01622	0,83675	50	561,04	26
2	**богъ**	1632	0,00466	1,15840	48	4,48	48
3	**нтн**	1546	0,00442	1,35581	43	567,40	25
4	**посълатн**	1408	0,00402	1,50897	34	1274,18	16
5	**зємла**	1375	0,00393	1,39811	37	592,37	24
6	**нзаславъ**	1254	0,00358	3,07925	1	3976,77	1
7	**володнмеръ/мнръ**	1243	0,00355	3,07915	2	3941,77	2
8	**велнкъ**	1226	0,00350	1,53935	30	1076,30	19
9	**рєщн**	1150	0,00329	1,17746	47	102,05	41
10	**къназь**	1146	0,00328	1,79778	19	2163,29	8
11	**дьнь**	1081	0,00309	1,42103	36	355,13	30
12	**сънъ**	1035	0,00296	1,37099	40	152,10	38

[10] Лемматизация осуществлена с помощью автоматического морфологического лемматизатора корпуса «Манускрипт» (вер. 5, http://manuscripts.ru/apex/f?p=104:LOGIN:17502242618790:::::), использующего электронный грамматической словарь. В ЛЛ лемматизировано 78% словоформ, в ЛИ – 81%, в ЛР – 67%.

№	Лемма	Абсолютное количество	Относительное количество	TF-ICTF'	Ранг	LL	Ранг
13	кнєвъ/кыєвъ	971	0,00277	3,07676	3	3076,25	3
14	мстнславъ	895	0,00256	3,07610	4	2834,43	4
15	городъ	889	0,00254	2,68598	13	2761,18	6
16	прнятн	872	0,00249	1,34794	44	25,02	45
17	понтн	786	0,00225	2,09342	17	1959,39	10
18	братъ	761	0,00217	1,71313	21	888,65	23
19	половьць	753	0,00215	3,07485	5	2382,67	7
20	дроужина	671	0,00192	2,15935	15	1717,65	13
21	всєволодъ	669	0,00191	3,07412	6	2115,46	9
22	наѹатн	663	0,00189	1,64088	25	518,63	27
23	отьць	659	0,00188	1,36858	41	0,13	50
24	съткорнтн	658	0,00188	1,33460	45	11,15	46
25	мъногъ	629	0,00180	1,39719	38	4,95	47
26	хотєтн	623	0,00178	1,52566	32	176,40	37
27	цьркы	601	0,00172	1,57690	28	263,08	33
28	ѧрославъ	587	0,00168	3,07340	7	1854,64	11
29	глаголатн	585	0,00167	1,21518	46	422,97	28
30	олегъ	581	0,00166	3,07334	8	1835,56	12
31	людн	543	0,00155	1,54936	29	143,40	39
32	градъ	528	0,00151	1,53462	31	105,67	40
33	протнвоу	523	0,00149	1,94594	18	951,80	21
34	лєто	510	0,00146	1,67379	24	348,69	31
35	ростнславъ	498	0,00142	3,07262	9	1571,60	14
36	възѧтн	492	0,00141	1,63889	26	251,32	34
37	мнръ	483	0,00138	1,37530	39	29,04	43
38	дроугъ	469	0,00134	1,45536	35	2,22	49
39	прнѧтн	469	0,00134	1,36596	42	47,33	42
40	статн	465	0,00133	1,74396	20	417,10	29
41	моужь	464	0,00133	1,61831	27	178,63	36
42	оубнтн	446	0,00127	1,70382	22	307,12	32
43	слъышатн	446	0,00127	1,51287	33	27,14	44
44	датн	439	0,00125	1,04122	49	2829,73	5
45	глєбъ	427	0,00122	3,07199	10	1345,86	15
46	вон	418	0,00119	2,20069	14	1015,86	20
47	даннлъ	412	0,00118	2,13615	16	930,77	22
48	братнѧ	379	0,00108	1,69367	23	191,50	35
49	роусь	378	0,00108	2,92089	12	1179,44	17
50	перєѧславль	371	0,00106	3,07150	11	1167,84	18

5.9. Четвертое обсуждение

В список входит несколько имен собственных – **нѯѧславъ**, **володнмеръ/володнмнръ**, **кнєвъ/кыєвъ**, **мстнславъ**, **всєволодъ**, **ѧрославъ**, **олегъ**, **ростнславъ**, **глєбъ**, **перєѧславль**, которые не встречаются в других коллекциях. Только в летописях встречается слово **половьць**. Единичны в текстах других жанров также **роусь** (2 раза) и **городъ** (6 раз). Понятно, что эти леммы получают наивысшее значение мер – от 2,68598 до 3,07915 по мере TF-ICTF' и от 1167,84 до 3941,77 – по LL.

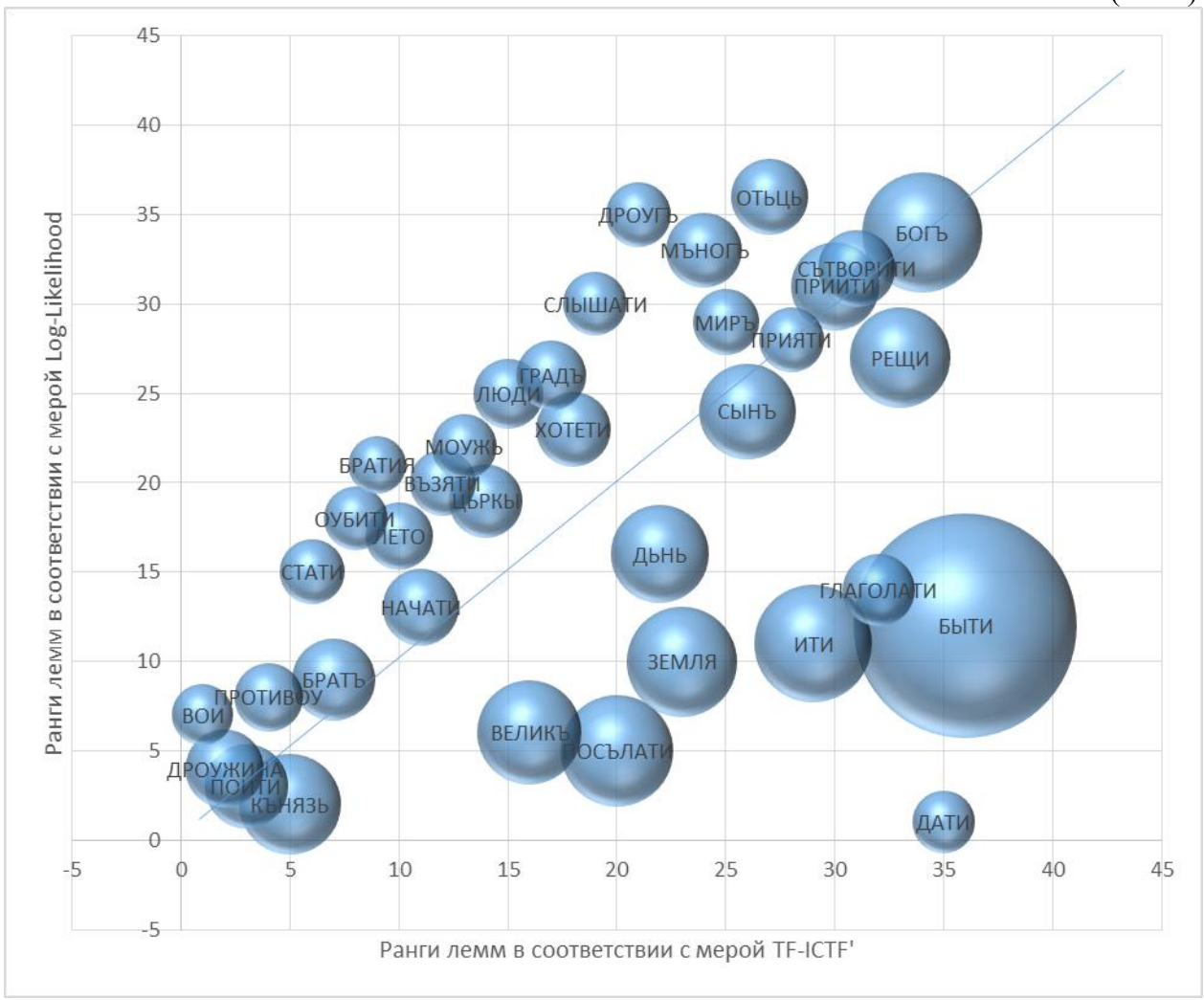
В соответствии с мерой LL статистическое значение практически всех 50-ти наиболее часто встречающихся слов свидетельствует о неслучайности их высокой

частоты в подкорпусе летописей: при значениях выше 15,31 это признается с вероятностью 99%. Лишь значения **сътворити**, **мъногъ**, **богъ**, **дроугъ**, **отьць**, которые в соответствии с мерой LL ниже и значительно ниже 15,31, не соответствуют выводу о неслучайности.

Мера TF-ICTF' максимально высоко оценивает имена собственные, встречающиеся единично в других коллекциях, – их ранги 1–13. В то же время мера LL к наиболее значимым относит не только имена собственные, которые занимают ранги от 1 до 18, но и некоторые другие леммы – **дати**, **кънязь**, **понти**, **дроужина**, **посълати**, ранги которых 5, 8, 10, 13 и 16 соответственно. Разные ранги в соответствии со значениями мер занимают и другие леммы.

Различия в оценке мерами лемм позволяют поставить вопрос о соотношении этих значений и о нахождении слов, которые мерами оцениваются одинаково высоко, одинаково низко и расходятся в оценке. Для выяснения этих соотношений представим данные в виде диаграммы, на которой по оси X даны ранги лемм в соответствии с мерой TF-ICTF', по оси Y – с мерой LL (количество лемм визуализируется величиной шара), см. диаграмму 3[11].

Диаграмма 3. Ранг меры TF-ICTF' & ранг меры LL & относительное количество 36-ти наиболее частотных знаменательных слов (лемм)



[11] В диаграмму не включены имена собственные, слова **половьць**, **роусь** и **городъ**, а также личное имя **данилъ**.

Одинаково высоко (ранги 1–10) обе меры оценивают леммы **понтн**, **дроужнна**, **кънѧзь**, **вон**, **протнвоу**, **братъ**[12]. Одинаково низко – **богъ**, **сътворнтн**, **прнятн** (ранги мер ниже 30), **прнѣатн** (ранги 28–28), **рєцн** (ранги 27–33), **отьць** (ранги 27–36).

Существенно различается оценка других лемм: леммы, оцениваемые высоко одной мерой, имеют низкий ранг в соответствии с другой, это хорошо видно по смещению лемм вправо и влево от линии с координатами 0–45. При этом леммы, встречающиеся максимально часто, высоко оцениваются мерой LL и низко мерой TF-ICTF' (ср. преобладание в нижней правой части диаграммы шаров больших размеров)[13].

Шесть выявленных статистически значимых слов (**понтн**, **дроужнна**, **кънѧзь**, **вон**, **протнвоу**, **братъ**) составляют ядро лексики летописей. Диаграмма позволяет увидеть и ближайшую периферию ядра – слова **наѵатн**, **статн** (ранги от 6 до 15), и более отдаленную – **оубнтн**, **лѣто**, **цьркъı**, **въѕатн**, **вєлнкъ**, **посълатн** (ранги от 5 до 20).

Слова, имеющие одинаково высокие статистические значения в соответствии с обеими мерами, следует признать не только уникальными для подкорпуса летописей (что справедливо в первую очередь в отношении имен собственных), но и ключевыми, отражающими наиболее актуальные темы текстов. Это темы КНЯЖЕСКОЙ ВЛАСТИ (**кънѧзь**, **дроужнна**, **вон**), РОДСТВА (**братъ**), ПОХОДОВ (**понтн**, **посълатн**), НАЧАЛА событий (**наѵатн**), ПРОТИВОСТОЯНИЯ и его результатов (**протнвоу**, **въѕатн**, **оубнтн**), УТВЕРЖДЕНИЯ событий (**статн**), ЦЕРКВИ (**цьркъı**), ВЕЛИЧИЯ (**вєлнкъ**) и ВРЕМЕНИ (**лѣто**).

6. Выводы

Создание исторических корпусов, в том числе славянских, увеличение их объема и расширение возможностей обработки, поиска и демонстрации данных позволяют перейти от этапа их использования для быстрого поиска текстовых иллюстраций и примеров, для сопоставления лексики и грамматики нескольких рукописей (подкорпусов) между собой, для установления точного количества лингвистических единиц к этапу анализа текстов нетрадиционными для историко-лингвистических исследований методами, в том числе – количественными и статистическими. Точная передача текста с помощью транскрипции, глубокая формально-структурная (аналитическая) разметка, включающая различного рода связи между аналитическими единицами (фрагментами), наличие лингвистических сведений о текстовых прецедентах, гибкая модель хранения данных, возможность точного и неточного

---

[12] Для сопоставления приведем леммы ядра (первые 10 рангов) других коллекций. Так, значимыми словами коллекции майских миней являются *песнь, славити, радовати, гласъ, верити, въпити, хрьстосъ, сила, вера*; миней на другие месяцы – *хрьстосъ, светъ, въпити, явити, славити, верити, песнь, божии*, стихирарей – *хрьстовъ, дьньсь, съпасти, милость, велии, доуша, слава, молити*; Евангелий – *ити, рещи, исоусъ, приити, имати, глаголати, оученикъ, дати, сътворити* [Баранов а]; Апостолов – *законъ, тело, доухъ, вера, божии* [Баранов 2017: 367]; Паренесисов – *помыслъ, страхъ, векъ, жити, хотети, братия* [Баранов б].

[13] Существенные различия в оценке леммы ДАТИ (ранги 35-1), скорее всего, вызваны включением в лемму совпадающего с аористной формой глагола союза ДА, который практически отсутствует в текстах других жанров и присвоением мерой LL высоких значений наиболее частотным лингвистическим единицам.

поиска позволяют получать разнообразные, в том числе и частотные, сведения о лингвистических единицах.

В основе историко-лингвистических исследований средневековых текстов практически всегда лежат данные о количестве как анализируемых, так и альтернативных единиц, а также о их общем количестве в рукописи или рукописях, из которых извлечены. Процедуры количественно-статистического анализа больших объемов данных похожи на ручную обработку относительно небольшого количества примеров: количество интересующих исследователя текстовых прецедентов сопоставляется с числом аналогичных или альтернативных форм на фоне всего объема лингвистических единиц корпуса. Подобная идентичность методик создает теоретические основания их применения для анализа исторических корпусов.

Сопоставление перечня наиболее частотных слов (форм) подкорпуса трех древнейших летописей (Лаврентьевской, Ипатьевской, Радзивилловской) с аналогичными списками слов семи разножанровых подкорпусов показало уникальность состава и порядка следования наиболее частотных слов в первом. В частности, позволило обнаружить отсутствие в нем наиболее частотных слов других подкорпусов, наличие большого количества пространственных предлогов и равнозначные по количеству форм совпадения с другими подкорпусами (раздел 5.3). В соответствии со статическими мерами TF-ICTF' и Log-Likelihood наибольшие значения в подкорпусе летописей имеют союз **а** и предлоги **съ** и **къ** (раздел 5.5).

Увеличение количества наиболее частотных слов до 25-ти дает аналогичные результаты: статистическим ядром списка являются союз **а** и предлоги **оу**, **о**, **съ**, **къ**, **по**, а также местоимение **тъ** и наречие **тако** (раздел 5.7).

Статистическая оценка 50-ти наиболее частотных знаменательных слов выделяет несколько имен собственных – антропонимы **нѫславъ**, **володимеръ/володимиръ**, **мстиславъ**, **всеволодъ**, **ѭрославъ**, **олегъ**, **ростиславъ**, **глѣбъ** и топонимы **кыевъ/кыѥвъ**, **переѭславль**, которые не встречаются в других коллекциях, а также слова **половьць**, **роусь городъ** (раздел 5.9).

Сопоставление оценок каждого слова списка (кроме собственных и трех уникальных нарицательных), полученных с помощью двух статистической мер, позволяет найти ядро перечня, в которое вошли леммы **понти**, **дроужина**, **къняжь**, **бон**, **противоу**, **братъ**, и его ближайшую и отдаленную периферию: **наѵати**, **стати**, **оубити**, **лѣто**, **цьркы**, **бъжати**, **велик**ъ, **посълати** (раздел 5.9).

Выделенные леммы с большой вероятностью можно признать ключевыми словами, отражающими наиболее актуальную тематику подкорпуса летописей: *княжеская власть*, *родство*, *походы*, *начало* событий, *противостояние*, его *результаты* и *утверждение* событий, *церковь*, *величие* и *время*.

В перспективе может быть осуществлен статистический анализ и каждой отдельной летописи, который должен позволить выявить тематику текста конкретной рукописи, а анализ фрагментов списков, выделенных, например, на основе разделения древнейших и позднейших их частей, – тематику каждой из них.

Благодарности

Статья написана при поддержке Российского фонда фундаментальных исследований (РФФИ) в рамках проекта «Лингвостатистический анализ однокомпонентных

и многокомпонентных лексических единиц исторического корпуса “Манускрипт”» (грант № 18-012-00463).

Источники

Лаврентьевская летопись, 1377 г., 173 л. (РНБ, F.п.IV.2). URL: http://manuscripts.ru/mns/main?p_text=32500902 (дата обращения: 08.04.2018).
Ипатьевская летопись, перв. пол. (сер.?) XV в., 307 л. (БАН, 16.4.4). URL: http://manuscripts.ru/mns/main?p_text=32151080 (дата обращения: 08.04.2018).
Радзивилловская летопись, кон. XV в., 245 л. (БАН, 34.5.30). URL: http://manuscripts.ru/mns/main?p_text=43296853 (дата обращения: 08.04.2018).

Основные машиночитаемые коллекции и корпуса средневековых славянских письменных памятников

Corpus Cyrillo-Methodianum Helsingiense: An Electronic Corpus of Old Church Slavonic Texts. URL: http://www.helsinki.fi/slaavilaiset/ccmh/ (дата обращения: 08.04.2018).
USC Parsed Corpus of Old South Slavic: The Historical Syntax of South Slavic. URL: http://www-bcf.usc.edu/~pancheva/ParsedCorpus.html (дата обращения: 08.04.2018).
Thesaurus Indogermanischer Text- und Sprachmaterialien: TITUS. URL: http://titus.uni-frankfurt.de/indexe.htm (дата обращения: 08.04.2018).
Regensburg Russian Diachronic Corpus. URL: http://rhssl1.uni-regensburg.de/SlavKo/korpus/rrudi-new (дата обращения: 08.04.2018).
Povest′ vremennyx let / D. Birnbaum (ed.), D. Ostrowski et al. (eds.). URL: http://pvl.obdurodon.org/ (дата обращения: 08.04.2018).
Old Russian Texts // Pragmatic Resources in Old Indo-European Languages. URL: http://foni.uio.no:3000 (дата обращения: 08.04.2018).
Old Russian Texts // TITUS. URL: http://titus.uni-frankfurt.de/indexe.htm (дата обращения: 08.04.2018).
Церковнославянский корпус // Национальный корпус русского языка. URL: http://www.ruscorpora.ru/search-orthlib.html (дата обращения: 08.04.2018).
Древнерусский корпус // Национальный корпус русского языка. URL: http://www.ruscorpora.ru/search-old_rus.html (дата обращения: 08.04.2018).
Старорусский корпус // Национальный корпус русского языка. URL: http://www.ruscorpora.ru/search-mid_rus.html (дата обращения: 08.04.2018).
Корпус берестяных грамот // Национальный корпус русского языка. URL: http://www.ruscorpora.ru/search-birchbark.html (дата обращения: 08.04.2018).
Корпус древнерусских берестяных грамот. URL: http://gramoty.ru/ (дата обращения: 08.04.2018).
Санкт-Петербургский корпус агиографических текстов. URL: http://project.phil.spbu.ru/scat/page.php?page=project (дата обращения: 08.04.2018).
Корпус «Манускрипт». URL: manuscripts.ru (дата обращения: 08.04.2018).

Перечни других электронных ресурсов, посвященных славянским средневековым рукописям – каталогов, библиотек сканированных списков и книг, изданий отдельных рукописей и др., см., например, на сайте «Общежитие: The World Wide Web portal for the study of Cyrillic and Glagolitic manuscripts and early printed books» (URL: http://www.obshtezhitie.net/) и на сайте «Inventory of Slavic, East European, and

Eurasian Digital Projects» (URL: https://www.library.illinois.edu/spx/inventory/projects.html).

Список литературы

Аллингтон и др. 2017 – Аллингтон, Д., Бруйетт, С., Голамбиа, Д. Неолиберальные инструменты (и архивы): политическая история цифровой гуманитаристики (Формы знания и критика их применения: новый век в «заботе о себе») [Электронный ресурс] // Гефтер. 25.01.2017. URL: http://gefter.ru/archive/20887 (дата обращения: 10.01.2018).

Баранов а – Опыт применения количественных и статистических методов для поиска значимых слов в историческом корпусе (на материале средневековых славянских гимнографических и евангельских кодексов) // Бонн, Германия (в печати).

Баранов б – Баранов В. А. Опыт статистического анализа славянского Паренесиса Ефрема Сирина (на материале электронной коллекции трех списков XIII–XIV вв. корпуса «Манускрипт») // Красноярск (в печати).

Баранов 2015 – Баранов В. А. Исторический корпус как цель и инструмент корпусной палеославистики // Scripta & e-Scripta: The Journal of Interdisciplinary Mediaeval Studies. Vol. 14–15. – Sofia: “Boyan Penev” Publishing Center ; Institute of Literature, BAS, 2015. – С. 39–62. – ISSN 1312-238X. URL: https://drive.google.com/file/d/0BwBejXXryRcROVQ4TnlpZFh6am8/view?usp=sharing (дата обращения: 08.04.2018).

Баранов 2017 – Баранов В. А. Статистически значимые слова как характеристика средневекового славянского текста (на материале коллекции Апостолов исторического корпуса «Манускрипт») // Гуманитарное образование и наука в техническом вузе [Электронный ресурс]: Сборник докладов Всероссийской научно-практической конференции с международным участием (Ижевск, 24–27 октября 2017 г.). – Ижевск: Изд-во ИжГТУ имени М. Т. Калашникова, 2017. – С. 359–369. URL: https://drive.google.com/file/d/1m0ZaYdM9EqEXvLjUz7emb8awkprdh_eB/view (дата обращения: 08.04.2018).

Ляшевская, Шаров 2009 – Ляшевская О. Н., Шаров С. А. Введение к частотному словарю современного русского языка // Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. С. 8. URL: http://dict.ruslang.ru/freq.pdf (дата обращения: 08.04.2018).

Kwok, K. L. 1995. A network approach to probabilistic information retrieval. In: ACM Transactions on Information Systems. Vol. 13. No 3. Pp. 324–353.

Rayson, Paul, and Roger Garside. 2000. Comparing corpora using frequency profiling // Proceedings of the Comparing Corpora Workshop at ACL 2000. Hong Kong, 2000. P. 1–6. URL: http://ucrel.lancs.ac.uk/people/paul/publications/rg_acl2000.pdf (дата обращения: 08.04.2018).

Robertson, S. 2004. Understanding inverse document frequency: on theoretical arguments for idf. In: Journal of Documentation. № 60. P. 503–520. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.2284&rep=rep1&type=pdf (дата обращения: 08.04.2018).

Roelleke, T., and Wang, J. 2006. A parallel derivation of probabilistic information retrieval models. In: Proceedings of the 29th Annual ACM SIGIR Conference on Re-

search and Development in Information Retrieval. Seattle, WA, S. Dumais, et al., Eds. ACM, New York. Pp. 107–114.

Roelleke, T. 2013. Information Retrieval Models: Foundations and Relationships. 141 p. URL: https://wiki.eecs.yorku.ca/course_archive/2014–15/F/4412/_media/ir_models.pdf (дата обращения: 08.04.2018).

Salton, G., and Yang, C. S. 1973. On the specification of term values in automatic indexing. In: Journal of Documentation. Vol. 29. Pp. 351–372.

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. In: Journal of Documentation. Vol. 28. Pp. 11–21.

Wu, H. C., Luk, R. W. P., Wong, K.F., and Kwok, K.L. 2008. Interpreting TF-IDF Term Weights as Making Relevance Decisions. In: ACM Transactions on Information Systems. Vol. 26. No. 3. Article 13. URL: https://www.scss.tcd.ie/khurshid.ahmad/Research/Sentiments/tfidf_relevance.pdf (дата обращения: 08.04.2018).

Приложение 1. Перечни 10-ти наиболее часто использующихся слов и словоформ в разножанровых подкорпусах

	Минеи на май[14]			*Другие минеи*			*Стихирари*			*Евангелия*			*Апостолы*[15]			*Паренесисы*			*Псалтыри*		
№	*W*[16]	*F*	*f*	*W*	*F*	*f*	*W*	*F*	*f*	*W*	*F*	*f*	*W*	*F*	*f*	*W*	*F*	*f*	*W*	*F*	*f*
1	н[17]	4296	0,043	н	8606	0,046	н	5377	0,051	н	40802	0,078	н	7341	0,056	н	17563	0,070	н	5151	0,069
2	са	3157	0,032	са	6070	0,032	са	3891	0,037	же	22041	0,042	же	6633	0,051	са	7786	0,031	са	2020	0,027
3	не	1802	0,018	не	3511	0,019	въ	2365	0,023	въ	15888	0,030	не	3686	0,028	не	7318	0,029	въ	1955	0,026
4	въ	1703	0,017	же	3369	0,018	же	1913	0,018	не	11565	0,022	въ	3837	0,029	въ	5306	0,021	же	1934	0,026
5	же	1634	0,016	въ	3061	0,016	не	1622	0,015	са	10597	0,020	са	3206	0,025	же	3977	0,016	не	1514	0,020
6	яко	1063	0,011	яко	2139	0,011	ѿ	1172	0,011	ѿ	8686	0,017	яко	1605	0,012	яко	3016	0,012	ѣко	1225	0,016
7	та	946	0,009	иси	2008	0,011	яко	1098	0,010	яко	6854	0,013	ѿ	1985	0,015	бо	2779	0,011	на	1024	0,014
8	на	900	0,009	пѣ	1724	0,009	на	1025	0,010	на	5874	0,011	о	1839	0,014	есть	2708	0,011	отъ	891	0,012
9	иси	841	0,008	та	1711	0,009	иси	901	0,009	емоу	4749	0,009	бо	1463	0,011	на	2605	0,010	бо	845	0,011
10	пѣ	812	0,008	на	1652	0,009	гла	873	0,008	его	4659	0,009	да	1273	0,010	ѿ	2388	0,010	ма	541	0,007

Приложение 2. Перечни 10-ти наиболее часто использующихся лемм знаменательных слов в разножанровых подкорпусах

	Минеи на май[18]		*Другие минеи*		*Стихирари*		*Евангелия*		*Апостолы*		*Паренесисы*		*Псалтыри*	
№	*L*	*F*	*L*	*F*	*L*	*F*	*L*	*F*	*L*	*F*	*L*	*F*	*L*	*F*
1	бытн	945	бытн	2985	бытн	1911	бытн	13201	бытн	3241	бытн	7016	бытн	1583
2	пѣснь	543	богъ	1175	богъ	993	рещн	4532	богъ	1103	богъ	1329	богъ	749
3	славнтн	405	хрьстосъ	775	молнтн	730	глаголатн	2972	глаголатн	563	глаголатн	1164	рещн	511
4	слово	385	слово	673	съпастн	559	нтн	2265	вѣра	528	нмѣтн	967	господь	445
5	богъ	367	божнн	589	доуша	478	оуученнкъ	1939	законъ	424	сътворнтн	755	земла	250
6	вѣрнтн	354	явнтн	690	сватъ	468	несоусъ	1850	божнн	359	дьнь	624	слово	214
7	хрьстосъ	349	свѣтъ	534	мнръ	416	прнтн	1846	нмѣтн	363	слово	604	людн	205
8	прнятн	329	вѣрнтн	532	дьнсь	412	вндѣтн	1676	еднъ	382	еднъ	597	сътворнтн	201
9	вѣра	296	молнтн	506	слава	383	сынъ	1689	несоусъ	289	рещн	575	вѣкъ	178
10	гласъ	301	родятн	510	вѣрнтн	358	еднъ	1697	доухъ	380	доуша	570	мнлость	158

[14] Общее количество словоформ в подкорпусах: списки служебной минеи на май – 99 613, служебные минеи на другие месяцы – 187 748, Стихирари – 104 905, Евангелия – 522 793, Апостолы – 130 779, Паренесисы – 249 124, Псалтыри – 74 343.

[15] Данные коллекции Апостолов отличаются от данных в предыдущих работах в связи с продолжающейся работой по разметке списков.

[16] *W* – словоформа, *L* – лемма, *F* – абсолютная частота, *f* – относительная частота.

[17] Учтены также графические варианты словоформ. В таблице указаны наиболее частотные варианты конкретной формы. Омонимия не снята.

[18] Общее количество лемматизированных форм в подкорпусах: списки служебной минеи на май – 60 571, служебные минеи на другие месяцы – 110 423, Стихирари – 85 326, Евангелия – 421 683, Апостолы – 99 749, Паренесисы –191 762, Псалтыри – 54 522.