Vilnius University Press

Contents lists available at Vilnius University Press

Slavistica Vilnensis
2019, vol. 64(2), pp. 10–30

ISSN 2351-6895 eISSN 2424-6115
DOI: https://doi.org/10.15388/SlavViln.2019.64(2).17

Статистически значимые слова четырех славянских служебных миней XI века

Виктор Аркадьевич Баранов

Ижевский государственный технический университет имени М. Т. Калашникова (Россия)
E-mail: victor.a.baranov@gmail.com

Аннотация. В статье описаны эксперименты по поиску ключевых слов в славянских служебных минеях XI века (РНБ, Соф. 202, РГАДА, ф. 381 № 84, 89, 91) с помощью трех статистических мер — Weirdness, TF*ICTF' и Log-Likelihood. В перечнях 30-ти наиболее частотных лемм каждого из списков (три другие рукописи выступают в качестве альтернативного подкорпуса) найдены статистически значимые слова и показана их различная оценка мерами. На основе сопоставления рангов лемм, ранжированных по значению мер, в каждом из списков выделены слова, наиболее значимые по оценке всех трех мер, а их сравнение с перечнями лемм, наиболее значимых в целом для коллекций гимнографических текстов корпуса “Манускрипт”, позволило найти ключевые слова каждой из рукописей. Установлены минимальные пороговые значения мер для выявления статистически значимых лингвистических единиц. Семантико-тематическая интерпретация ключевых слов позволяет увидеть их принадлежность к основным понятиям христианской церкви, охватывающим идею о *свете*, *чистоте*, *святости*, *радости*, *прославлении*, *спасении*, *покое, истине* и ее конкретизацию в каждой из рукописей: МП — *победа над страданием и смертью*, *святость*, *восхваление*, *радость*, МС — *чистота*, *Бог*, МО — *истина*, *покой*, МН — *восхваление*, *истина*, *спасение*.
Ключевые слова: славянские служебные минеи; ключевые слова; лингвистическая статистика

Statistically significant words of four Service Menaia of the 11th century

Abstract. The work deals with experiments on search of keywords in Slavonic Service Menaia of the 11th century by means of three statistic measures — Weirdness, TF*ICTF' and Log-Likelihood. Statistically significant words are found in the lists of 30 the most frequent lemmas of each copy (three other manuscripts play the role of the alternative subcorpus) and their different assessments by the respective measures are shown. The words that are the most significant due to the assessments done by the three measures on the basis of comparison of the ranks of the lemmas ranked by the values of measures are found in each list. Their comparison with the lists of lemmas that are the most significant on the whole for the collections of hymnographic texts of the corpus “Manuscript” helps to find the keywords in each

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (РФФИ) в рамках проекта «Лингвостатистический анализ однокомпонентных и многокомпонентных лексических единиц исторического корпуса “Манускрипт”» (проект № 18-012-00463).

Received: 3/6/2019. **Accepted:** 25/12/2019

of the manuscripts. Minimal threshold values of the measures are established for revelation of statistically significant linguistic units. Semantic-thematic interpretation of the keywords gives a possibility of seeing their relationship with the main notions of the Christian church and its realization in each manuscript: Menaion of May — *defeat of suffering and death*, *holiness*, *praise*, *joy*, Menaion of September — *purity*, *God*, Menaion of October — *truth*, *peace*, Menaion of November — *praise*, *truth*, *salvation*.
Keywords: Slavonic Service Menaia; keywords; linguistic statistics

---

Keturių XI amžiaus slaviškų liturginių minėjų statistiškai reikšmingi žodžiai

Anotacija. Straipsnyje aprašomi eksperimentai, skirti raktažodžių paieškai XI amžiaus liturginių minėjų tekstuose naudojant tris statistikos matus: Weirdness, TF*ICTF' ir Log-Likelihood. Trisdešimties dažniausiai naudojamų lemų sąrašuose nustatyti statistiškai reikšmingi žodžiai ir parodytas skirtingas šių žodžių vertinimas matais. Kiekviename sąraše nustatyti aukščiausiai visais trimis matais vertinami žodžiai. Lyginant šiuos žodžius su himnografinių tekstų korpuso „Manuscript“ kolekcijų reikšmingų žodžių sąrašu nustatyti kiekvieno analizuojamo rankraščio raktažodžiai. Rastos minimalios slenkstinės matų reikšmės statistiškai reikšmingiems lingvistiniams vienetams nustatyti. Raktažodžių semantinė ir teminė interpretacija leidžia priskirti juos pagrindinėms krikščionių bažnyčios sąvokoms ir konkretinti kiekvienam iš rankraščių: gegužės mėn. minėja – *kančios ir mirties nugalėjimas*, *šventumas*, *garbinimas, džiaugsmas*, rugsėjo mėn. minėja – *skaistumas*, *Dievas*, spalio mėn. minėja – tiesa, ramybė, lapkričio mėn. minėja – *garbinimas, tiesa, išvadavimas*.
Reikšminiai žodžiai: slavų liturginės minėjos; raktažodžiai; lingvistinė statistika

---

1. Постановка задачи

Количественные и статистические методы давно и успешно применяются для анализа текстов и создания компьютерных систем обработки информации и анализа документов.

В основе компьютерных систем лежат большие массивы текстовых данных, как существующих в электронной форме изначально, так и переведенных в машиночитаемую форму из печатной.

Одним из хорошо разработанных направлений в области статистической лингвистики является автоматический поиск ключевых слов. Базовым принципом является сопоставление лексики анализируемого документа с лексикой альтернативного (контрастного) корпуса, где частота конкретного слова принимается за среднюю частоту. Временная, жанровая, гендерная и иная разметка документов корпуса позволяет установить различия в частотности лингвистических единиц различных групп текстов и соотнести статистические значения единиц анализируемого документа с соответствующими единицами той или иной группы документов, а на основании сведений о группе сделать вывод не только о жанре, времени создания, авторе текста, но и о формально-количественных, дистрибутивных, сочетаемостных, лексико-тематических, лексико-семантических особенностях документа.

За рамками статистического анализа до сих пор оставались средневековые славянские тексты, что было связано с отсутствием как достаточно больших машиночитаемых коллекций средневековых произведений, так

и специализированных компьютерных систем для подготовки, хранения, разметки, обработки, анализа и демонстрации таких коллекций.

В последние десять лет были подготовлены такие ресурсы, отличающиеся графико-орфографической точностью транскрипции, специализированной мета- и лингвистической разметкой, наличием дополнительных компонентов, обеспечивающих автоматическую лемматизацию, особых параметров запросов и визуализации выборок и др.[1]

Один из таких ресурсов — информационно-аналитическая система "Манускрипт", содержащая тексты средневековых славянских рукописей X–XV вв. (корпус "Манускрипт" — manuscripts.ru) и произведений М. В. Ломоносова (корпус Ломоносова — lomonosov.pro). Система функционирует на платформе СУБД Oracle, имеет специальный редактор для создания транскрипций и разметки текстов и морфологический автоматический анализатор, использующий правила унификации текстовых прецедентов и нормализованные словарные формы грамматической базы данных, предоставляет пользователям несколько модулей поиска и демонстрации текстов, конкордансов, перечней словоформ, лемм, n-грамм, параллельных корпусов и др.

Объем (более 3,5 млн словоформ), мета- и аналитическая разметка для формирования подкорпусов, разнообразные параметры запросов и выдачи результатов позволяют использовать средневековый корпус для поиска данных при решении разнообразных традиционных историко-лингвистических задач. Особым направлением применения данных корпуса является их анализ с помощью количественных и статистических методов.

В настоящее время сведения о количественных характеристиках выборки можно получить с помощью нескольких запросных форм. Модуль n-грамм дает возможность осуществить оценку словосочетаний с помощью статистических мер, а также позволяет построить указатель одиночных словоформ или лемм, сортированный по частоте встречаемости. Наличие сведений о количестве каждой лингвистической единицы в документе или подкорпусе позволяет использовать такой перечень в качестве первичных данных при статистическом анализе форм текста (коллекции, подкорпуса).

Одной из задач, для решения которых необходимо привлечение количественных данных, является автоматическое извлечение значимых для документа или группы документов ключевых лингвистических единиц (форм слов или слов и их сочетаний), то есть наиболее соответствующих тематике текста компонентов. Кроме того, такой перечень статистически значимых словоформ или слов, а также состав и порядок следования единиц в ранжированном по частоте встречаемости списке, их относительное количество и некоторые другие характеристики могут рассма-

[1] См. список электронных коллекций и корпусов в конце работы.

триваться как индивидуальные особенности документа, отличающие его от других документов корпуса.

2. Цели, методы и материал работы

В основе целей данной работы — предположение о том, что статистические методы могут быть эффективными и результативными при применении их в отношении средневековых текстов, входящих в корпуса относительно небольшого объема.

Для эксперимента по поиску ключевых слов средневековых текстов были выбраны четыре древнейшие славянские служебные минеи XI века — Путятина минея (РНБ, Соф. 202) (далее — МП), минеи на сентябрь, октябрь и ноябрь (РГАДА, ф. 381 № 84, 89, 91) (далее — МС, МО, МН) и три статистические меры, которые используются для поиска ключевых слов в документах.

2.1. Статические меры

С целью автоматического нахождения в документе терминов предметной области, а также для оценки близости текста документа текстам других документов используются различные статистические методы.

2.1.1. Метод Weirdness

Метод Weirdness (“Странность”) относится к группе, в которой применяются данные “контрастного”, “альтернативного” подкорпуса [Ahmad, Gillam, Tostevin 1999] и вычисляется отношение между относительными частотами слов в анализируемом документе и в альтернативной коллекции:

$$Weirdness_w = \frac{W_s/T_s}{W_g/T_g},$$ где

“W_s — частотность слова в коллекции предметной области,
T_s — совокупная частотность слов в коллекции предметной области,
W_g — частотность слова в контрастной коллекции (или W_g + 1 для предотвращения деления на 0 при отсутствии слова в контрастной коллекции [Gillam, Tariq, Ahmad 2005]),
T_g — совокупная частотность слов в контрастной коллекции” [Клышинский, Кочеткова 2014: 368].

Статистически значимыми словами признаются те, значение меры которых выше и значительно выше 1,0.

Метод характеризуется как один из традиционно применяемых для извлечения часто встречающихся в документе терминов и используется, например, при работе с небольшими коллекциями [Бессмертный, Чуцяо,

Пенюй 2016, 1097, 1098], [Бессмертный, Нугуманова, Мансурова, Байбурин 2017, 81, 83, 85].

*2.1.2. Метод TF*ICTF'*

Мера TF*ICTF' представляет собой один из многочисленных вариантов известной меры TF*IDF (term frequency — inverse document frequency) [Sparck 1972], [Salton, Yang 1973], [Roelleke 2013], [Robertson 2004], где значимость слова определяется на основе его относительной частоты в документе и количества документов корпуса, в которых слово встречается:

$TF*IDF = \frac{f}{F} * log \frac{D}{d}$, где

f — количество анализируемых слов (term) в документе,
F — количество всех слов в документе,
D — количество документов в корпусе,
d — количество документов корпуса, в которых встречается анализируемое слово.

В варианте TF*ICTF (term frequency — inverse collection term-frequency) [Kwok 1995], [Roelleke, Wang 2006], [Wu et al 2008, 17–18] учитывается не количество текстов, в которых встречается анализируемое слово, а количество слов в них:

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

f_d — количество анализируемых слов (term) в документе,
F_d — количество всех слов в документе,
F_D — общее количество слов в корпусе,
f_D — количество анализируемых слов во всех документах корпуса.

Различные модификации меры позволяют уменьшить смещение значения меры в сторону наиболее частотных слов (см., например, [Roelleke 2013]):

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

где $F_D - F_d$ — объем корпуса без объема документа, в которую входит анализируемое слово,
$f_D - f_d$ — количество слов во всех коллекциях, кроме его количества в анализируемом документе.

2.1.3. Метод Log-Likelihood

Мера *Log-Likelihood* (коэффициент правдоподобия, показатель сходства) (см., например, [Rayson, Garside 2000: 3], [Ляшевская, Шаров 2009, XI–

XII]), используя данные о частотности слова в анализируемом документе и объеме текста, а также сведения о встречаемости слова в альтернативном подкорпусе и об объеме последнего, позволяет, как и две предыдущие, найти в документе тематически значимые слова:

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

где *a* — абсолютное количество анализируемого слова в документе,
b — абсолютное количество анализируемой единицы в альтернативных документах,
c — объем анализируемого документа,
b — объем альтернативных документов.

2.2. Цели работы и результаты предыдущих экспериментов

2.2.1. Цели работы:

- выявление в каждой из четырех славянских миней XI века слов, которые можно считать ключевыми, а их состав — индивидуальной характеристикой документа,
- сопоставление оценки слов статистическими мерами для получения непротиворечивых результатов,
- выяснение возможностей каждой из трех статистических мер для нахождения значимых слов средневекового документа.

2.2.2. Некоторые результаты предыдущих экспериментов

Сравнение подкорпуса служебных миней XI–XIV вв. с текстами других жанров[2] в [Баранов 2019] позволило показать индивидуальность состава и порядка следования наиболее частых служебных и знаменательных слов в каждом из подкорпусов, выявить статистически значимые служебные слова и леммы в каждом из подкорпусов, обнаружить значительную близость гимнографических подкорпусов друг другу и их контрастность по отношению к подкорпусу евангельских списков.

Анализ с помощью мер TF*ICTF' и Log-Likelihood позволил выявить статистически значимые леммы из первых 16-ти наиболее частотных в каждом из подкорпусов:

TF*ICTF':

- в минеях на май (в порядке уменьшения значения меры) — **радоватн**, **славнтн**, **въпнтн**, **снла**, **глась**, **пѣснь**,

[2] Коллекция майских миней XI–XIII вв.: количество рукописей — 4, отрывков — 3; объем — 99 613 словоформы; коллекция миней на сентябрь, октябрь, ноябрь, февраль-август, апрель XI–XIV вв.: количество рукописей — 6; объем — 187 748 словоформ; коллекция стихирарей XII (XIV, XV) вв.: количество рукописей — 4, отрывков — 1; объем — 104 905 словоформ; коллекция евангелий XI–XIV вв.: количество рукописей — 9, отрывков — 2; объем — 522 793 словоформы; общий объем четырех коллекций — 915 059 словоформ.

– в минеях на другие месяцы — **въпнтн**, **слабнтн**, **хрьстосъ**, **пѣснь**, **доуша**, **свѣтъ**;

Log-Likelihood:

– в минеях на май — **пѣснь**, **бъıтн**, **слабнтн**, **радобатн**, **вѣрнтн**, **гласъ**,
– в минеях на другие месяцы — **хрьстосъ**, **богъ**, **ѭбнтн**, **бъıтн**, **свѣтъ**, **вѣрнтн**.

Сопоставление рангов лемм в перечнях позволило выделить слова, имеющие наиболее высокое значение в соответствии с обеими мерами. К таким словам относятся:

– в майских минеях — **пѣснь**, **слабнтн**, **радобатн** (ранги 1–5), **гласъ**, **вѣрнтн**, **хрьстосъ**, **въпнтн**, **снла**, **вѣра** (ранги 6–10),
– в минеях на другие месяцы — **хрьстосъ** (ранги 1–2), **свѣтъ**, **ѭбнтн**, **вѣрнтн**, **божнн**, **въпнтн**, **слабнтн**, **пѣснь** (ранги 5–10).

При анализе текстов одного жанра использование в качестве альтернативных текстов других жанров выявляет слова, контрастные для коллекции или жанра, но не для каждого отдельного документа.

С целью выявления значимых для каждого отдельного текста слов в настоящей работе альтернативным подкорпусом для каждой из рукописей стали тексты того же жанра — трех других миней.

2.3. Характеристика данных

Были рассмотрены 30 наиболее частых лемм знаменательных слов каждой из четырех древнейших славянских миней XI века.

2.3.1. Объем списков

МП — 23648 словоформ, 13890 словоформ приведены к лемме[3] (58,74%);

МС — 41534 словоформы, 24484 словоформы приведены к лемме (58,95%);

МО — 31221 словоформа, 17941 словоформа приведена к лемме (57,47%);

МН — 37999 словоформ, 21516 словоформ приведены к лемме (56,62%).

[3] Лемматизация списков осуществлена автоматически с помощью морфологического анализатора корпуса «Манускрипт» (http://manuscripts.ru/apex/f?p=104:LOGIN:1089214070511:::::).

2.3.2. Наиболее частотные леммы списков

Перечни 30-ти наиболее частых лемм[4] каждой из рукописей приведены в таблице 1 (приложение 1). Списки содержат достаточно большое количество совпадающих слов: **бꙑтн**, **богъ**, **слово**, **хрьстосъ**, **ѭвнтн**, **славнтн**, **прнѭтн**, **вѣрнтн**, **въпнтн**, **свѣтъ**, **вѣра**, **божнн**, **молнтн**, **нꙑнѣ**, **снла**, **веселнтн**, **роднтн**, **мнръ**, **слава** (всего — 19 лемм).

Одновременно в каждый из списков входят и слова, повторяющиеся в двух других перечнях или только в одном из них, например, в ПМ — **пѣснь**, **гласъ**, **ꙁемлѧ**, **пѣтн**, в МС — **господь**, **доуша**, **кднньъ**, **ꙁемлѧ**, **нстнна**,**уоудо**, **пѣснь**, **просвѣтнтн**, **доухъ**, в МО — **просвѣтнтн**, **господь**, **доухъ**, **кднньъ**, **нстнна**, **доуша**, **пѣтн**, **уоудо**, **оумъ**, в МН — **доуша**, **пѣтн**, **просвѣтнтн**, **кднньъ**, **господь**, **оумъ**, **нстнна**, и леммы, не повторяющиеся в альтернативном подкорпусе: в ПМ — **вѣньць**, **въснѭтн**, **радоватн**, **страсть**, **блаженъ**, **любꙑ**, **ꙋстьнъ**, в МС — **ангелъ**, **унстъ**, в МО — **разоумъ**, **вндѣтн**, в МН — **цьркꙑ**, **съпастн**, **прѣунстъ**, **матн**. В случае отсутствия анализируемой леммы среди 30-ти форм трех других списков статистическая оценка осуществляется на основе количества соответствующих лемм за пределами этих перечней.

3. Анализ

3.1. Статистические значения лемм

В таблице 2 раздела 3.4 приведены статистические значения и соответствующие им ранги лемм ПМ[5].

Наибольшие значения имеют те леммы текста, которые в трех других списках встречаются значительно реже. Так, очень частотная лемма **пѣснь**, встретившаяся 269 раз в ПМ и лишь 95 раз в трех других минеях, имеет значения 1,55 (TF*ICTF'), 13,26 (Weirdness) и 553,64 (LL), лемма **гласъ** (138 vs 62) — 1,64, 10,42 и 255,92 соответственно. Кроме того в ПМ очень высокие значения имеет и находящаяся всего лишь на 25-м месте по частоте использования лемма **радоватн** (50 раз в ПМ, только 1 раз в других списках) — 2,53 (TF*ICTF'), 234,17 (Weirdness) и 164,30 (LL).

В МС из наиболее частотных наибольшие значения мер имеют леммы **роднтн**, **господь**, из лемм второй половины списка — **ангелъ** и

[4] Выборка осуществлена с помощью модуля n-грамм корпуса (http://manuscripts.ru/mns/cred_ngr.stat), позволяющего выбрать не только многокомпонентные сочетания, но и отдельные словоформы или леммы и расположить их в порядке частоты встречаемости.

[5] Аналогичные вычисления были сделаны для наиболее частых лемм МС, МО И МН.

ѵнїтъ. В МО наибольшие значения имеют леммы, находящиеся ближе к концу списков, — **ꙗазꙋмъ**, **дꙋхъ**, в МН — лемма **цьрькы** из середины перечня.

При этом видно, что лемма, имеющая наибольший вес в соответствии с одной мерой, может не иметь самого высокого ранга в соответствии с другой. Например, в МО в соответствии с мерами TF*ICTF' и Weirdness наибольшее значение имеет лемма **ꙗазꙋмъ**, в соответствии с LL — **бꙑтн**.

Понятно, что различия в оценке лемм связаны как с количеством конкретной леммы в анализируемом тексте и в альтернативном подкорпусе, так и с различиями в оценке лингвистических единиц каждой из статистических мер. В целом видно, что LL высоко оценивает и наиболее, и наименее частотные леммы, мера TF*ICTF' — менее частотные, мера

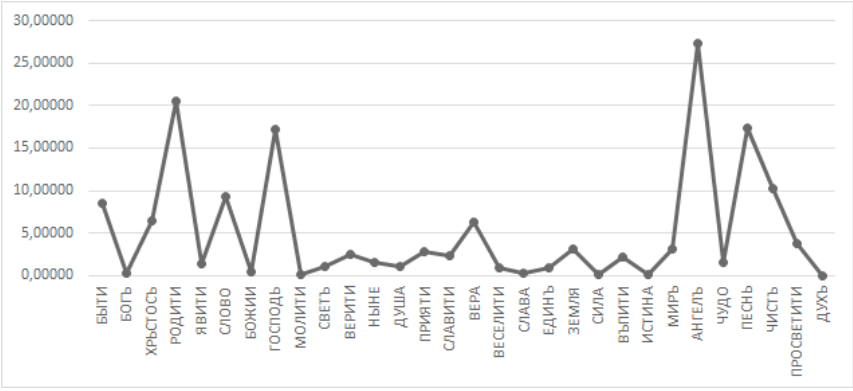

Диаграмма 1. *Значения лемм МС в соответствии с мерой Log-Likelihood*

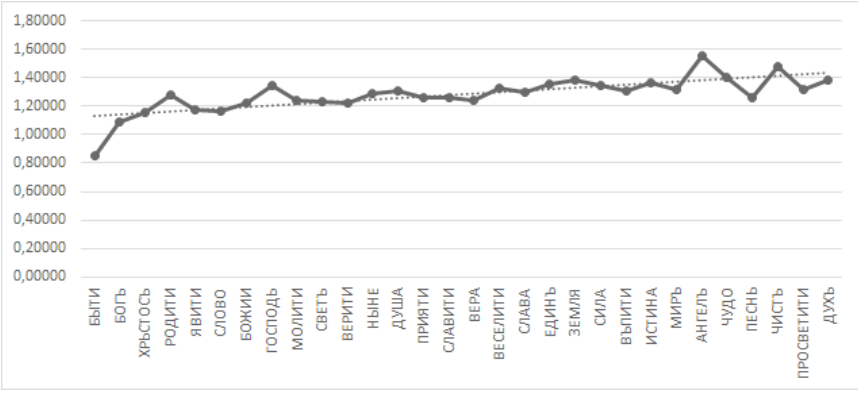

Диаграмма 2. *Значения лемм МС в соответствии с мерой TF*ICTF'*

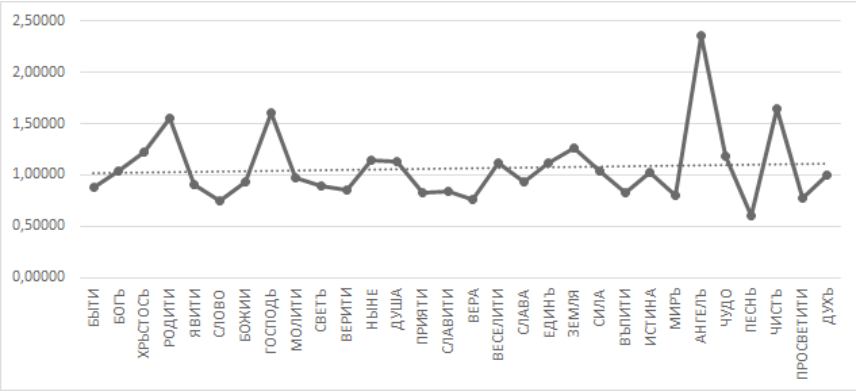

Диаграмма 3. *Значения лемм МС в соответствии с мерой Weirdness*

Weirdness — выделяет наиболее и наименее частотные, присваивая наибольшие значения вторым (см. диаграммы 1–3[6]).

Попарное сопоставление значений мер (см. диаграммы 4–6) позволяет увидеть различное отношение статистических мер к текстовым

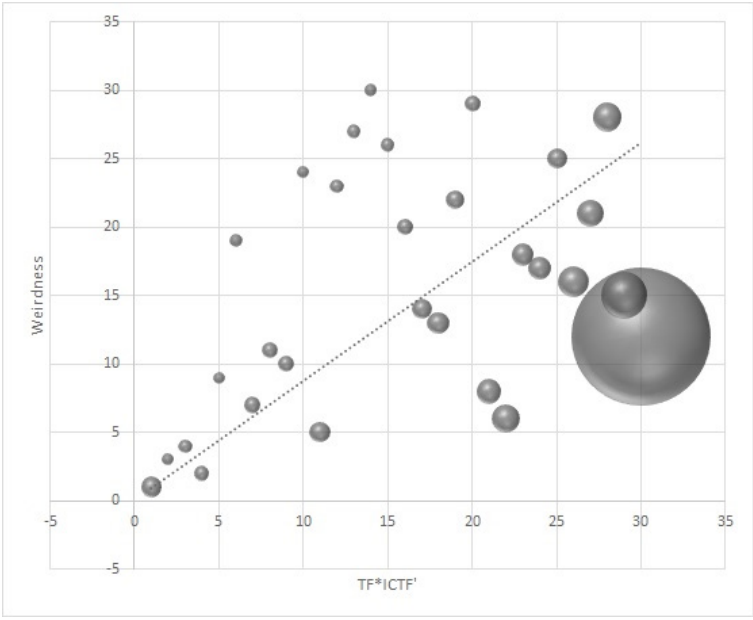

Диаграмма 4. *Распределение лемм по рангам в соответствии с мерами TF*ICTF' и Weirdness*[7]

[6] Диаграммы построены с помощью программы Excel MS Office.

[7] По осям диаграммы приведены ранги двух статистических мер; диаметр шара соответствует относительному количеству леммы.

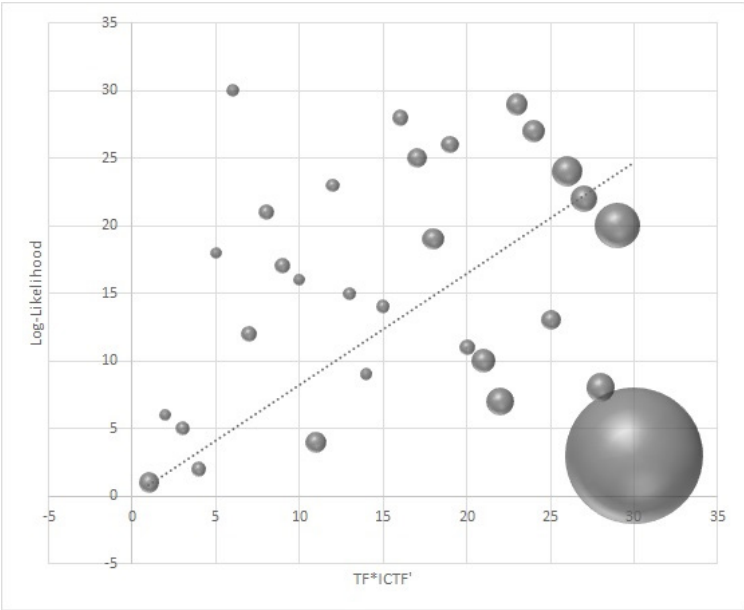

Диаграмма 5. *Распределение лемм по рангам в соответствии с мерами TF*ICTF' и Log-Likelihood*

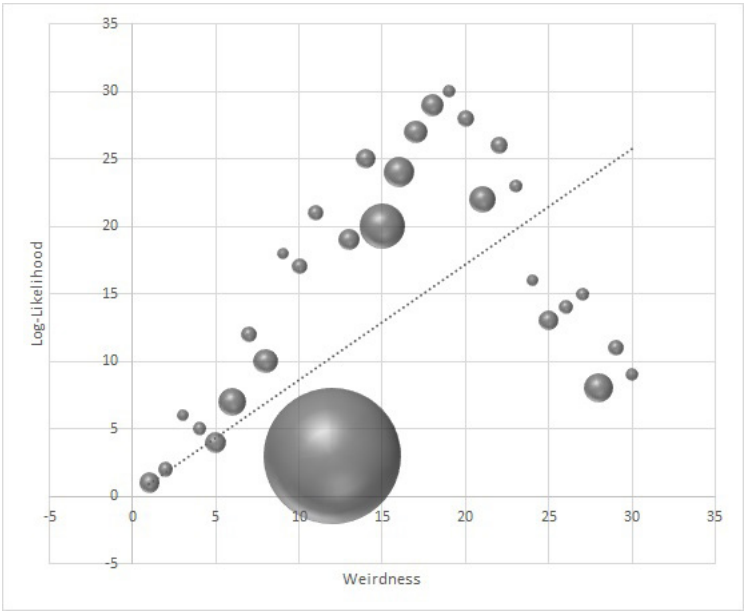

Диаграмма 6. *Распределение лемм по рангам в соответствии с мерами Weirdness и Log-Likelihood*

единицам с разной частотой встречаемости. Так, диаграмма 4 показывает, что мера Weirdness более частые леммы оценивает более высоко, чем мера TF*ICTF' (крупные шары занимают место правее и ниже линии тренда). На диаграмме 5 крупные шары, находящиеся в правой части диаграммы и отсутствующие в левой, позволяют говорить о том, что мера LL оценивает наиболее частые леммы выше, чем мера TF*ICTF'. Диаграмма 6, демонстрирующая соотношение рангов в соответствии с мерами Weirdness и Log-Likelihood, показывает, что леммы со средней частотой первая мера оценивает более высоко, чем вторая (см. находящиеся левее и выше линии тренда значительное количество лемм из середины списка).

3.2. Ранги лемм

Ранжирование лемм в соответствии с их весом по каждой из мер позволяет сопоставить леммы между собой и выделить леммы, которые всеми тремя мерами оцениваются наиболее высоко, низко, а также различно и противоположно.

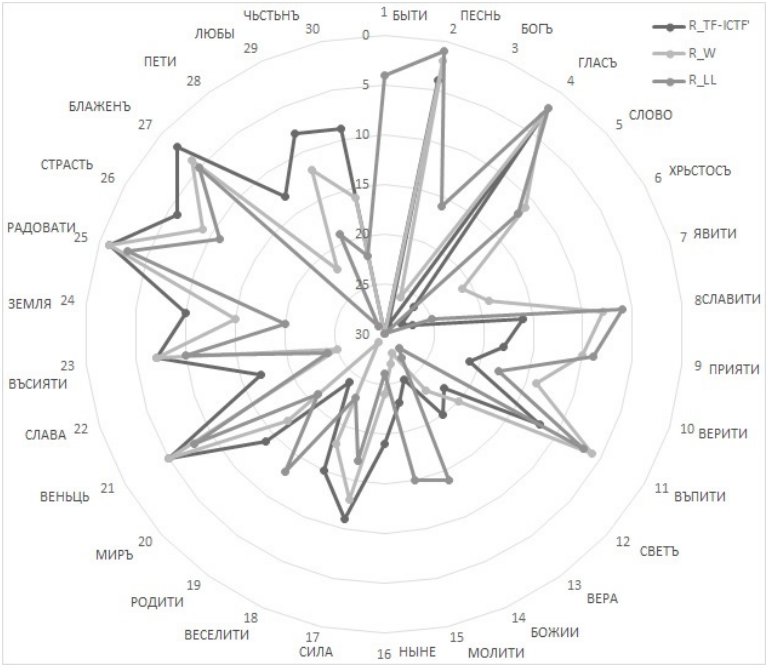

Диаграмма 7. *Ранги 30-ти наиболее частотных лемм ПМ*[8]

[8] Леммы располагаются по часовой стрелке от наиболее частотных к менее частотным, значения вершин диаграмм соответствуют рангам лемм, наибольшее значение имеют леммы, вершины которых находятся у края окружности.

Использование лепестковых диаграмм для сопоставления рангов дает возможность найти леммы, имеющие одинаково высокий ранг в соответствии со всеми мерами (см. диаграммы 7–10). Идентичная или близкая оценка леммы всеми тремя мерами позволяет сделать вывод, что именно эта лемма, являясь одной из 30-ти наиболее частотных, одновременно может оцениваться как максимально контрастная по отношению к трем другим спискам.

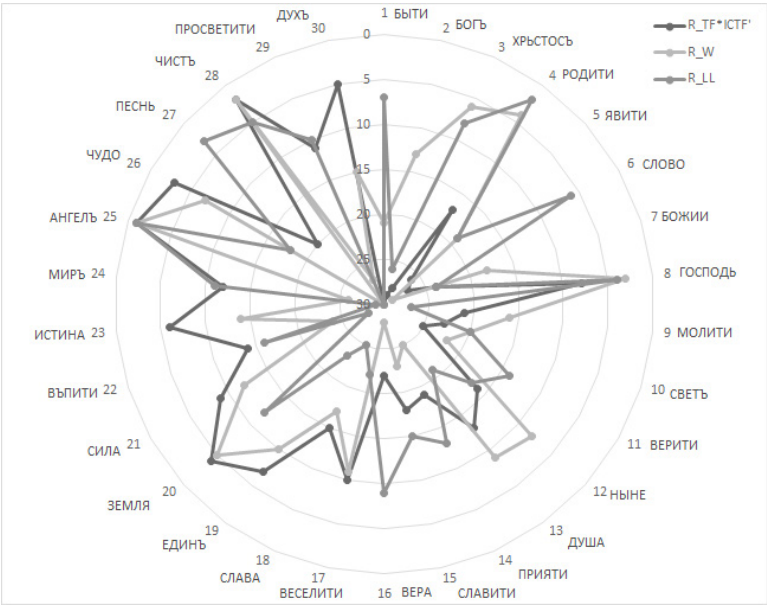

Диаграмма 8. *Ранги 30-ти наиболее частотных лемм МС*

ПМ[9]. В пределах 1–5 рангов всех трех мер находятся леммы **пѣснь** (1-4 — максимальные ранги леммы), **гласъ** (2-3), **радовати** (1-3), **блаженъ** (2-5), 5–10 рангов — **вѣньць** (5-8), **въсньати** (7-10), 6–12 — **въпнти** (6-12), **страсть** (6-11). Другие леммы оцениваются мерами или существенно различно (**богъ**, **слово**, **славнти**, **роднти** и др.), или противоположно (**бъıти**).

МС. Ранги 1–5 занимают леммы **ангелъ** (1-1), **чистъ** (2-5), 3–8 — **господь** (4-8). Все остальные оцениваются мерами или существенно различно (**роднти**, **доуша**, **земла** и др.), или противоположно (**слово**, **мнръ** и др.).

[9] Показаны леммы, имеющие ранги не ниже 12 позиции и разницу между рангами не более 6 пунктов. Значения 12 и 6 подобраны экспериментально.

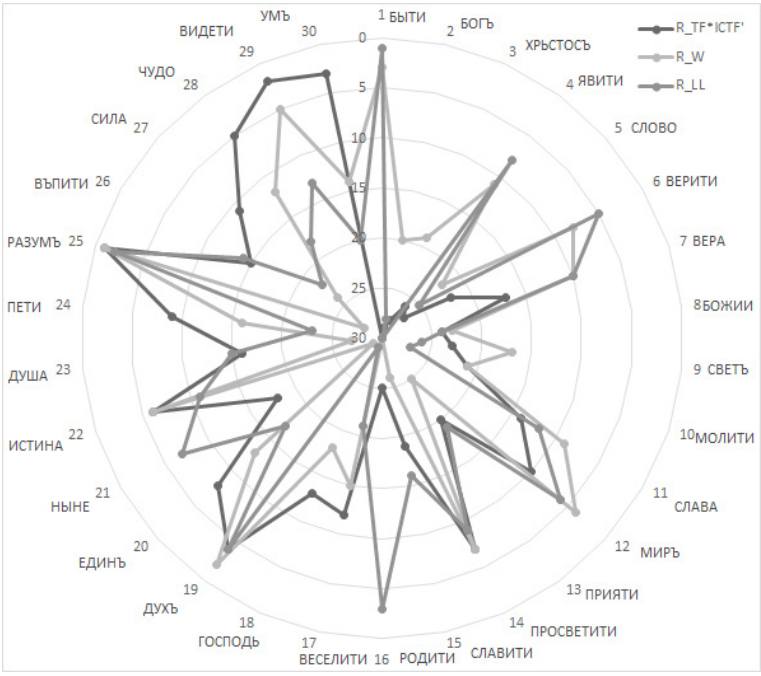

Диаграмма 9. *Ранги 30-ти наиболее частотных лемм МО*

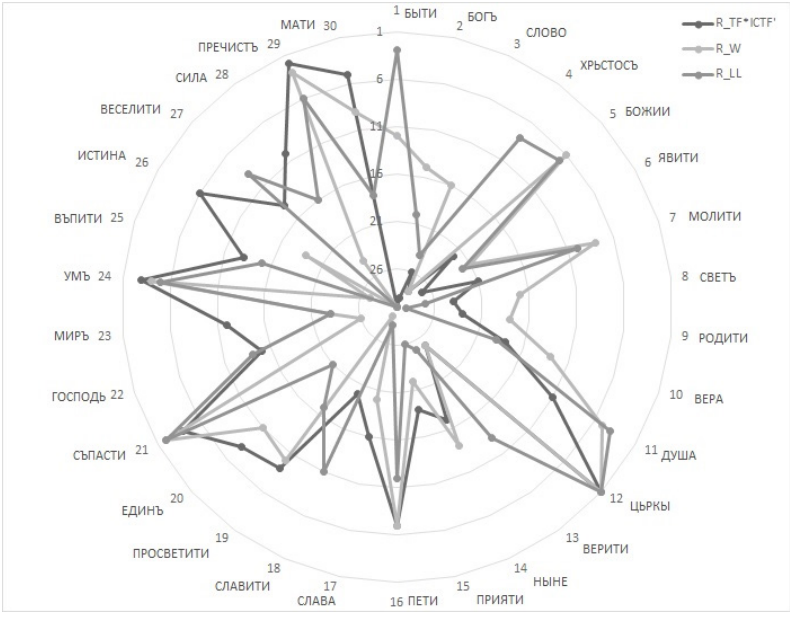

Диаграмма 10. *Ранги 30-ти наиболее частотных лемм МН*

МО. Ранги 1–4 имеют леммы **доухъ** (2-4), **разоумъ** (1-2), 4–11 — **мнръ** (4-10), **просвѣтнтн** (7-9), **нстнна** (6-11). Существенно различно оцениваются **веселнтн**, **доуша**, **въпнтн** и др., противоположно — **бытн**, **вѣрнтн**, **роднтн**, **оумъ** и др.

МН. Ранг 1 имеет лемма **цьркъı** (1-1), ранги 2–6 — **съпастн** (2-4), **оумъ** (3-5), **преүнстъ** (2-6), **пѣтн** (7-12). Существенно различающиеся ранги имеют **молнтн**, **господь** и др., совершенно различные — **бытн**, **божнн** и нек. др.

Наибольшее значение для выявления ключевых слов имеют леммы, находящиеся на позициях высоких рангов всех трех статистических мер.

Сопоставление значимых для ПМ лемм с наиболее значимыми леммами майских миней в целом (см. раздел 2.2.2) позволяет увидеть как совпадение результатов в пределах первых 12 рангов **пѣснь**, **глась**[10], **радовати**, **въпнтн**, так и дополнительные леммы **блаженъ**, **вѣньць**, **въснѧтн**, **страсть**, которые можно считать ключевыми.

Аналогичное сравнение значимых лемм всех миней и МС, МО и МН в отдельности показывает уникальность лемм, занимающих первые 6–12 рангов, для каждой из рукописей: ключевые леммы МС — **ангелъ**, **үнстъ**, **господь**, МО — **доухъ**, **разоумъ**, **мнръ**, **просвѣтнтн**, **нстнна**, МН — **цьркъı**, **съпастн**, **оумъ**, **преүнстъ**, **пѣтн**[11].

Статистические значения лемм в текстах соотносимы друг с другом:

МП: TF*ICTF' — 2,5–1,4, Weirdness — 4,6–1,8, LL — 553,6–10,3;

МС: TF*ICTF' — 1,6–1,4, Weirdness — 2,4–1,6, LL — 27,3–10,3;

МО: TF*ICTF' — 1,5–1,4, Weirdness — 1,8–1,3, LL — 12,9–2,9;

МН: TF*ICTF' — 1,5–1,4, Weirdness — 6,2–1,3, LL — 75,2–3,6.

В соответствии с мерой TF*ICTF' ключевыми являются леммы, статистическое значение которых выше 1,4, Weirdness — выше 1,3, LL — выше 2,9.

3.3. Семантико-тематическая интерпретация ключевых слов

Все выявленные наиболее значимые для текстов слова — 12 существительных, 5 глаголов и 2 прилагательных — являются символами христи-

[10] Оба слова используются в заголовках чтений.

[11] Все леммы, отнесенные к ключевым, имеют статистические значения в соответствии с мерой Weirdness выше и значительно выше 1,0. Леммы **пѣснь**, **глась**, **радовати**, **въпнтн**, **блаженъ**, **вѣньць**, **въснѧтн**, **цьркъı**, **ангелъ**, **господь** имеют статистические значения в соответствии с мерой Log-Likelihood выше величины 15,31, что с вероятностью 99% свидетельствует о неслучайности частоты встречаемости леммы в составе документа/коллекции; за пределами списка ключевых слов в соответствии с мерой LL остались **бытн** (МП, МО), **славнтн** (МП), **прнѧтн** (МП), **роднтн** (МС), не соответствующие критериям — ранг выше 12, разница рангов не более 6 пунктов.

анства. Их семантика соотносима с основными понятиями христианской *церкви*, и каждое из них может считаться центральным для соответствующих семантических полей. К ключевым словам этой группы относится и лемма **вѣньць**, которая является в тексте ПМ символом победы над смертью: **побѣдьнъ вѣньць**, **неистлѣньнъ вѣньць**, **вѣньць нетлѣньнъ**, **вѣньць правьдьнъ**, **вѣньць неоувадаемъ**, **пресвѣтлъ вѣньць** и др. Лемма **страсть** употребляется в контекстах почитания страданий, победы над страданием: **дьнесь страсти поунтаемъ**, **поунтажщимъ страсти твоѭ**, **страсти въсхвалѭштимі**, **хвалѭще страсти твоѭ**, **оумрьтвн страсти плътьскыѭ**, **прогонншн страстн лютыі** и др., что также соотносится с семантикой других ключевых слов[12].

При этом каждый из текстов имеет свою, более конкретную семантику ключевых слов: МП — *победа над страданием и смертью*, *святость*, *восхваление*, *радость*, МС — *чистота*, *Бог*, МО — *истина*, *покой*, МН — *восхваление*, *истина*, *спасение*.

3.4. Количественные и статистические данные о 30-ти наиболее частых леммах служебных миней XI века

Таблица 1. *Перечни 30-ти наиболее частых лемм четырех служебных миней XI века*[13]

№	*Путятина минея*		*Сентябрьская минея*		*Октябрьская минея*		*Ноябрьская минея*	
	W	*F*	*W*	*F*	*W*	*F*	*W*	*F*
1	бытн	308	бытн	762	бытн	786	бытн	835
2	пѣснь	269	богъ	290	богъ	212	богъ	275
3	богъ	140	хрьстосъ	253	хрьстосъ	166	слово	177
4	гласъ	138	родитн	185	ѭвитн	158	хрьстосъ	172
5	слово	130	ѭвитн	172	слово	135	божнн	162
6	хрьстосъ	125	слово	149	вѣрнтн	129	ѭвитн	159
7	ѭвитн	110	божнн	141	вѣра	111	молитн	143
8	славнтн	108	господь	136	божнн	109	свѣтъ	135
9	прнѭтн	99	молитн	133	свѣтъ	108	родитн	130
10	вѣрнтн	97	свѣтъ	131	молитн	101	вѣра	127
11	въпитн	90	вѣрнтн	130	слава	93	доуша	120
12	свѣтъ	83	нъінѣ	126	мнръ	85	цьркъı	120
13	вѣра	69	доуша	113	прнѭтн	84	вѣрнтн	116

[12] Ср. ключевые леммы с более общими и менее конкретизированными значениями *силы*, *света*, *бога*, *веры*, *прославления*, *радости*, выявленные при сопоставлении миней с текстами других жанров (см. разд. 2.2.2).

[13] *W* — лемма, *F* — абсолютное количество в тексте списка.

№	*Путятина минея*		*Сентябрьская минея*		*Октябрьская минея*		*Ноябрьская минея*	
	W	*F*	*W*	*F*	*W*	*F*	*W*	*F*
14	божнн	67	прнѣтн	106	просвѣтнтн	80	нъінѣ	113
15	молнтн	62	слабнтн	105	славнтн	79	прнѣтн	103
16	нъінѣ	62	вѣра	104	роднтн	77	пѣтн	93
17	снла	60	веселнтн	103	веселнтн	76	слава	93
18	веселнтн	59	слава	101	господь	76	славнтн	92
19	роднтн	59	ѥдннъ	94	доуχъ	76	просвѣтнтн	91
20	мнръ	59	ζемла	90	ѥднъ	72	ѥднъ	88
21	вѣньць	57	снла	88	нъінѣ	71	съпастн	85
22	слава	56	въпнтн	85	нстнна	70	господь	79
23	въснѣтн	55	нстнна	79	доуша	68	мнръ	78
24	ζемла	51	мнръ	79	пѣтн	68	оумъ	77
25	радобатн	50	ангелъ	78	разоумъ	66	въпнтн	76
26	страсть	50	чоудо	78	въпнтн	62	нстнна	71
27	блажєнъ	49	пѣснь	77	снла	61	веселнтн	70
28	пѣтн	47	чнстъ	75	чоудо	58	снла	69
29	любъі	46	просвѣтнтн	74	вндѣтн	56	преунстъ	68
30	чьстьнъ	44	доуχъ	73	оумъ	53	матн	65

Таблица 2. *Перечень 30-ти наиболее частых лемм МП, их статистические значения и ранги*[14]

№	*W*	*F*	*f*	R_{TF}	*TF*ICTF'*	R_W	*Weirdness*	R_{LL}	*LL*
1	бъітн	308	0,01302	30	0,84448	30	0,60533	4	78,12716
2	пѣснь	269	0,01138	4	1,55076	2	13,26153	1	553,63837
3	богъ	140	0,00592	29	1,08335	26	0,84386	16	3,55493
4	глась	138	0,00584	3	1,63547	3	10,42443	2	255,92235
5	слово	130	0,00550	26	1,19687	11	1,32071	12	7,43911
6	χрьстосъ	125	0,00529	28	1,14239	21	0,99058	30	0,00927
7	ѭвнтн	110	0,00465	27	1,18300	19	1,05353	25	0,24155
8	славнтн	108	0,00457	16	1,30767	8	1,83265	6	25,84553
9	прнѣтн	99	0,00419	18	1,29414	10	1,58246	9	14,38677
10	вѣрнтн	97	0,00410	21	1,24023	14	1,21145	18	2,72914
11	въпнтн	90	0,00381	12	1,35316	6	1,89018	7	23,51298
12	свѣтъ	83	0,00351	22	1,24008	20	1,03937	28	0,10047
13	вѣра	69	0,00292	20	1,25883	23	0,94490	27	0,18672
14	божнн	67	0,00283	25	1,21817	28	0,76163	14	4,55908
15	молнтн	62	0,00262	23	1,23724	27	0,77022	15	3,86102
16	нъінѣ	62	0,00262	19	1,27985	24	0,93669	26	0,22424
17	снла	60	0,00254	11	1,35638	13	1,28902	17	2,88704
18	веселнтн	59	0,00249	15	1,32738	18	1,10973	23	0,50604
19	роднтн	59	0,00249	24	1,22859	29	0,70491	13	6,82950

[14] *№* — ранг леммы, *W* — лемма, *F* — абсолютное количество, *f* — относительное количество, R_{TF} — ранг в соответствии с мерой TF-ICTF', *TF-ICTF'* — значение меры, R_W — ранг в соответствии с мерой *Weirdness*, *Weirdness* — значение меры, R_{LL} — ранг в соответствии с мерой Log-Likelihood, *Log-Likelihood* — значение меры.

№	*W*	*F*	*f*	R_{TF}	*TF*ICTF'*	R_W	*Weirdness*	R_{LL}	*LL*
20	мнръ	59	0,00249	14	1,33359	17	1,14183	21	0,81211
21	вѣньць	57	0,00241	5	1,47042	5	2,06943	8	18,77229
22	слава	56	0,00237	17	1,29630	25	0,91384	24	0,38810
23	въснѭтн	55	0,00233	7	1,45562	7	1,86659	10	13,87406
24	ѯемла	51	0,00216	10	1,37681	15	1,20634	20	1,37462
25	радовати	50	0,00211	1	2,52751	1	234,17202	3	164,29881
26	страсть	50	0,00211	6	1,46171	9	1,74755	11	10,34706
27	блажєнъ	49	0,00207	2	1,67616	4	4,58977	5	52,39990
28	пѣти	47	0,00199	13	1,34304	22	0,95291	29	0,09184
29	любы	46	0,00195	8	1,41357	12	1,29005	19	2,22665
30	ѹстьнъ	44	0,00186	9	1,39714	16	1,14484	22	0,62928

4. Выводы

Использование статистических мер для выявления в средневековых текстах ключевых слов позволило увидеть общую тональность жанра и тематику отдельного текста.

Сопоставление с текстами других жанров дало возможность выявить ключевые слова наиболее общей семантики анализируемого жанра (в славянских служебных минеях XI–XIV веков это *сила*, *свет*, *Бог*, *вера*, *прославление*, *радость*), сопоставление текстов одного жанра между собой позволило уточнить и конкретизировать семантику жанра (в служебных минеях XI века это *победа*, *свет*, *чистота*, *святость*, *радость*, *прославление*, *спасение*, *покой, истина, вечность*) и определить ключевые слова каждого из списков (МП — *победа над страданием и смертью*, *святость*, *восхваление*, *радость*, МС — *чистота*, *бог*, МО — *истина*, *покой*, МН — *восхваление*, *истина*, *спасение*).

Эмпирически найденная условная граница между значимыми и незначимыми леммами позволила установить значения каждой из статистических мер, выше которых леммы можно считать ключевыми: для TF*ICTF' — это 1,4, для Weirdness — 1,3, для Log-Likelihood — 2,9.

Таким образом, данные относительно небольшого по объему исторического корпуса “Манускрипт” могут быть использованы для постановки нетрадиционных для историко-лингвистических исследований задач и для их эффективного и результативного решения количественными и статистическими методами.

Источники

МП — Минея служебная на май (Путятина минея) XI в. (РНБ, Соф. 202), 135 л. URL: http://manuscripts.ru/mns/main?p_text=16723192 (дата обращения: 20.03.2019).

МС — Минея служебная на сентябрь 1095–1096 г. (РГАДА, ф. 381 № 84), 176 л. URL: http://manuscripts.ru/mns/main?p_text=32821351 (дата обращения: 20.03.2019).

МО — Минея служебная на октябрь 1096 г. (РГАДА, ф. 381 № 89), 127 л. URL: http://manuscripts.ru/mns/main?p_text=33118322 (дата обращения: 20.03.2019).

МН — Минея служебная на ноябрь 1097 г. (РГАДА, ф.381 № 91), 174 л. URL: http://manuscripts.ru/mns/main?p_text=33347210 (дата обращения: 20.03.2019).

Древнерусский корпус // Национальный корпус русского языка. URL: http://www.ruscorpora.ru/search-old_rus.html

Корпус берестяных грамот // Национальный корпус русского языка. URL: http://www.ruscorpora.ru/search-birchbark.html

Корпус древнерусских берестяных грамот. URL: http://gramoty.ru/

Корпус “Манускрипт”. URL: manuscripts.ru

Корпус М. В. Ломоносова. URL: lomonosov.pro

Санкт-Петербургский корпус агиографических текстов. URL: http://project.phil.spbu.ru/scat/page.php?page=project

Старорусский корпус // Национальный корпус русского языка. URL: http://www.ruscorpora.ru/search-mid_rus.html

Церковнославянский корпус // Национальный корпус русского языка. URL: http://www.ruscorpora.ru/search-orthlib.html

Old Russian Texts // Pragmatic Resources in Old Indo-European Languages. URL: http://foni.uio.no:3000

Old Russian Texts // TITUS. URL: http://titus.uni-frankfurt.de/indexe.htm

Povest’ vremennyx let / D. Birnbaum (ed.), D. Ostrowski et al. (eds.). URL: http://pvl.obdurodon.org/

Regensburg Russian Diachronic Corpus. URL: http://rhssl1.uni-regensburg.de/SlavKo/korpus/rrudi-new

Литература

БАРАНОВ, В. А., 2019. Опыт применения количественных и статистических методов для поиска значимых слов в историческом корпусе (на материале средневековых славянских гимнографических и евангельских кодексов). *Studia Hymnographica.* Band II, eds.Hans Rothe und Claudia Schnell. Verlag Ferdinand Schöningh. (=Patristica Slavica, Band 24, eds. Hans Rothe. Abhandlungen der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste, Bd. 131), 149–201.

БЕССМЕРТНЫЙ, И. А., ЮЙ ЧУЦЯО, МА ПЕНЮЙ, 2016. Статистический метод извлечения терминов из китайских текстов без сегментации фраз. *Научно-технический вестник информационных технологий, механики и оптики*, 16, № 6, 1096–1102. doi: 10.17586/2226-1494-2016-16-6-1096-1102. URL: http://ntv.ifmo.ru/file/article/16157.pdf.

БЕССМЕРТНЫЙ, И. А., НУГУМАНОВА, А. Б., МАНСУРОВА, М. Е., БАЙБУРИН, Е. М., 2017. Метод контрастного извлечения редких терминов из текстов на естественном языке. *Научно-технический вестник информационных технологий, механики и оптики*, 17, № 1, 81–91. doi: 10.17586/2226-1494-2017-17-1-81-91. URL: http://ntv.ifmo.ru/file/article/16383.pdf.

КЛЫШИНСКИЙ, Э. С., КОЧЕТКОВА, Н. А., 2014. Метод извлечения технических терминов с использованием меры странности, *Новые информационные техноло-*

гии в автоматизированных системах, 17, 365–370. URL: https://elibrary.ru/download/elibrary_21527004_14693581.pdf.

ЛЯШЕВСКАЯ, О. Н., ШАРОВ, С. А., 2009. Введение к частотному словарю современного русского языка. In: Ляшевская О. Н., Шаров С. А. (авт.). *Частотный словарь современного русского языка (на материалах Национального корпуса русского языка)*. Москва, V–XXII.

AHMAD, K., GILLAM, L., TOSTEVIN, L., 1999. University of surrey participation in TREC8: weirdness indexing for logical document extrapolation and retrieval (WILDER). In: *Proc. 8th Text Retrieval Conference TREC*. Gaithersburg, USA, 717–724.

GILLAM, L., TARIQ, M., AHMAD, K., 2005. Terminology and the construction of ontology. *Terminology*, V, 11, 1, 55–81.

KWOK, K. L., 1995. A network approach to probabilistic information retrieval. *ACM Transactions on Information Systems*, 13, No 3, 324–353.

RAYSON, P., GARSIDE, R., 2000. Comparing corpora using frequency profiling. In: *Proceedings of the Comparing Corpora Workshop at ACL 2000*. Hong Kong, 1–6. URL: http://ucrel.lancs.ac.uk/people/paul/publications/rg_acl2000.pdf.

ROBERTSON, S., 2004. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60, 503–520. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.2284&rep=rep1&type=pdf.

ROELLEKE, T., WANG, J., 2006. A parallel derivation of probabilistic information retrieval models. In: Dumais S., et al. (eds.), *Proceedings of the 29th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. Seattle, New York, 107–114.

ROELLEKE, T., 2013. *Information Retrieval Models: Foundations and Relationships*. URL: https://wiki.eecs.yorku.ca/course_archive/2014–15/F/4412/_media/ir_models.pdf.

SALTON, G., YANG, C. S., 1973. On the specification of term values in automatic indexing. *Journal of Documentation*, 29, 351–372.

SPARCK JONES, K., 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28, 11–21.

WU, H. C., LUK, R. W. P., WONG, K. F., KWOK, K. L., 2008. Interpreting TF-IDF Term Weights as Making Relevance Decisions. *ACM Transactions on Information Systems*, 26, No 3, Article 13. URL: https://www.scss.tcd.ie/khurshid.ahmad/Research/Sentiments/tfidf_relevance.pdf.

Bibliography (Transliteration)

BARANOV, V. A., 2019. Opyt primeneniya kolichestvennyh i statisticheskih metodov dlya poiska znachimyh slov v istoricheskom korpuse (na materiale srednevekovyh slavyanskih gimnograficheskih i evangel'skih kodeksov). *Studia Hymnographica.* Band II, eds. Hans Rothe und Claudia Schnell. Verlag Ferdinand Schöningh. (= Patristica Slavica. Band 24, eds. Hans Rothe. Abhandlungen der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste, Bd. 131), 149–201.

BESSMERTNYJ, I. A., YUJ CHUCYAO, MA PENYUJ, 2016. Statisticheskij metod izvlecheniya terminov iz kitajskih tekstov bez segmentacii fraz. *Nauchno-tekhnicheskij vestnik informacionnyh tekhnologij, mekhaniki i optiki*, 16, № 6, 1096–1102. doi: 10.17586/2226-1494-2016-16-6-1096-1102. URL: http://ntv.ifmo.ru/file/article/16157.pdf.

BESSMERTNYJ, I. A., NUGUMANOVA, A. B., MANSUROVA, M. E., BAJBURIN, E. M., 2017. Metod kontrastnogo izvlecheniya redkih terminov iz tekstov na estestvennom yazyke, *Nauchno-tekhnicheskij vestnik informacionnyh tekhnologij, mekhaniki i optiki*, 17, № 1, 81–91. doi: 10.17586/2226-1494-2017-17-1-81-91. URL: http://ntv.ifmo.ru/file/article/16383.pdf.

KLYSHINSKIJ, E. S., KOCHETKOVA, N. A., 2014. Metod izvlecheniya tekhnicheskih terminov s ispol'zovaniem mery strannosti. *Novye informacionnye tekhnologii v avtomatizirovannyh sistemah*, 17, 365–370. URL: https://elibrary.ru/download/elibrary_21527004_14693581.pdf.

LYASHEVSKAYA, O. N., SHAROV, S. A., 2009. Vvedenie k chastotnomu slovaryu sovremennogo russkogo yazyka. In: Lyashevskaya O. N., SHarov S. A. (avt.). *CHastotnyj slovar' sovremennogo russkogo yazyka (na materialah Nacional'nogo korpusa russkogo yazyka)*. Moskva, V–XXII.

AHMAD, K., GILLAM, L., TOSTEVIN, L., 1999. University of surrey participation in TREC8: weirdness indexing for logical document extrapolation and retrieval (WILDER). In: *Proc. 8th Text Retrieval Conference TREC*. Gaithersburg, USA, 717–724.

GILLAM, L., TARIQ, M., AHMAD, K., 2005. Terminology and the construction of ontology, *Terminology*, V. 11, N 1, 55–81.

KWOK, K. L., 1995. A network approach to probabilistic information retrieval. *ACM Transactions on Information Systems*, 13. No 3. 324–353.

RAYSON, P., GARSIDE, R., 2000. Comparing corpora using frequency profiling. In: *Proceedings of the Comparing Corpora Workshop at ACL 2000*. Hong Kong, 1–6. URL: http://ucrel.lancs.ac.uk/people/paul/publications/rg_acl2000.pdf.

ROBERTSON, S., 2004. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60, 503–520. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.2284&rep=rep1&type=pdf.

ROELLEKE, T., WANG, J., 2006. A parallel derivation of probabilistic information retrieval models. In: Dumais S., et al. (eds.). *Proceedings of the 29th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. Seattle, New York, 107–114.

ROELLEKE, T., 2013. *Information Retrieval Models: Foundations and Relationships*. URL: https://wiki.eecs.yorku.ca/course_archive/2014–15/F/4412/_media/ir_models.pdf.

SALTON, G., YANG, C. S., 1973. On the specification of term values in automatic indexing. *Journal of Documentation*, 29, 351–372.

SPARCK JONES, K., 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28, 11–21.

WU, H. C., LUK, R. W. P., WONG, K. F., KWOK, K. L., 2008. Interpreting TF-IDF Term Weights as Making Relevance Decisions. *ACM Transactions on Information Systems*, 26, No. 3, Article 13. URL: https://www.scss.tcd.ie/khurshid.ahmad/Research/Sentiments/tfidf_relevance.pdf.

Баранов Виктор Аркадьевич, д-р филол. наук, профессор, заведующий кафедрой лингвистики Ижевского государственного технического университета имени М. Т. Калашникова

Baranov Victor Arkadievich, Doctor of Philology, Full Professor, Head of the Department of Linguistics, Izhevsk State Technical University named after M.T. Kalashnikov

Baranov Viktor Arkadjevič, filologijos mokslų daktaras, profesorius, Iževsko valstybinio M. T. Kalašnikovo technikos universiteto Lingvistikos katedros vedėjas