ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

УДК 519.2:801.82(045)
DOI 10.22213/2410-9304-2018-4-65-74

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ БИГРАММ РУССКИХ ЕВАНГЕЛЬСКИХ СПИСКОВ XI–XIV ВЕКОВ *

В. А. Баранов, доктор филологических наук, профессор, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия

С помощью корреляционного анализа биграмм рассмотрена степень близости русских списков Евангелий XI–XIV веков разного типа – полных апракосов, кратких апракосов, тетра – друг другу, а также степень близости Пантелеймонова Евангелия XII века (полный апракос) каждому из типов. Анализу подвергнуты перечни биграмм с наибольшим значением T-score, компонентами которых являются леммы; объем выборок из каждого подкорпуса – 300 элементов. Для выявления близости перечней к рангам биграмм применена непараметрическая статистика r-Спирмена, к значениям в соответствии со статистической мерой T-score – статистика r-Пирсона.

Полученные результаты позволяют сделать выводы о наличии корреляционной связи между сопоставляемыми массивами биграмм, которая имеет высокую статистическую вероятность, а также о достаточно существенной степени корреляции, которая характеризуется или как умеренная, или как заметная. Оценка силы связи между подкорпусами позволяет говорить и о различиях в степени близости сопоставляемых массивов биграмм. В соответствии с ранговой корреляцией r-Спирмена наибольшую близость обнаруживают подкорпус полных и подкорпус кратких апракосов, а также Пантелеймоново Евангелие и полные апракосы, наименьшую – краткие апракосы и тетр, а также Пантелеймоново Евангелие и тетр (или краткие апракосы). В соответствии с корреляцией r-Пирсона наибольшая близость выявлена между полными апракосами и тетром, наименьшая – между полными и краткими апракосами. Отношения Пантелеймонова Евангелия с тетром и краткими апракосами аналогичны отношениям с ними полных апракосов.

Ключевые слова: лингвистическая статистика, биграмма, Евангелие, русские списки, полный апракос, краткий апракос, тетр, Пантелеймоново Евангелие.

1. Статистика и средневековые славянские Евангелия

Состав, структура и язык древнейших славянских Евангелий исследовались многократно (см., например, [1; 2; 3]). Работами А. А. Алексеева, его коллег и учеников в России положено начало анализу Евангелий с помощью статистических методов [4] [1] и доказана эффективность применения последних для решения задач кластеризации евангельских списков, для установления их текстологической и лингвистической близости.

В работе [5] дан количественно-статистический анализ биграмм и триграмм Пантелеймонова Евангелия (РНБ, Соф. 1, XII в.) в сопоставлении с соответствующими *n*-граммами коллекции других русских евангельских рукописей XI–XIV веков, электронные машиночитаемые копии которых размещены в историческом корпусе «Манускрипт» (manuscripts.ru) [2]. Сравнение дву- и трехкомпонентных сочетаний, имеющих наибольшее значение в соответствии со статистической мерой T-score, показало, что совпадения и различия между ними обусловлены не лингвистическими, а, скорее всего, структурными особенностями рукописей: количеством повторений одних и тех же контекстов (чтений) в Пантелеймоновом Евангелии и подкорпусе евангельских списков.

2. Евангельские рукописи различного типа как цель статистического анализа

Цели данной работы: а) выявление степени близости евангельских рукописей различного типа друг другу; б) установление места полноапракосного Пантелеймонова Евангелия XII века (РНБ, Соф. 1) (далее – ПЕ) относительно каждого из подкорпусов.

Основанием для формулирования целей является гипотеза о том, что различия в составе чтений между типами рукописей – кратким (далее – КА) и полным (далее – ПА) апракосами, а также тетром (далее – Т) – может оказывать влияние на количественные и статистические характеристики наиболее частотных биграмм – формально выделенных двукомпонентных сочетаний лингвистических единиц.

* 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).

[1] Среди последних работ ученых этой научной школы – [6; 7].

[2] О корпусе см., например, [8].

В связи со сложными, до конца не установленными текстологическими и лингвистическими отношениями между славянскими рукописями разных типов, между рукописями внутри одного типа, а также между частями одной рукописи (см., например, [9; 10; 11]) необходимо выяснить степень зависимости некоторых общих характеристик лингвистических единиц от типа рукописи. Степень близости, выраженная количественно, могла бы стать некоторой средней величиной для статистического сопоставления, например, отдельных рукописей между собой, текстов разных редакций, списков разных классов внутри типа в дальнейшем.

В работе представлены результаты анализа статистических характеристик биграмм трех коллекций евангельских рукописей разного типа – полных апракосов, кратких апракосов и тетра [3] – в сопоставлении друг с другом и Пантелеймонова Евангелия (полный апракос) в сопоставлении с характеристиками биграмм каждого из подкорпусов.

В отличие от других работ, в которых для установления степени близости евангельских списков друг другу анализируется некоторый общий фрагмент, в данной работе выборка данных осуществлена из всего объема рукописей.

3. Обоснование корреляционного метода

Для выяснения степени близости подкорпусов (текстов) в качестве анализируемых массивов могут быть выбраны аналогичные и сопоставимые перечни лингвистических единиц, извлеченные из подкорпусов и упорядоченные по одному и тому же критерию, например, по частоте встречаемости.

Анализ таких перечней (массивов) биграмм может быть осуществлен с помощью корреляционного анализа, который, как известно, позволяет выявить связь двух или нескольких рядов величин, при которой изменение значений одного массива данных соответствует или не соответствует изменению значений другого массива. Корреляция свидетельствует о связи явлений, одно из которых может быть причиной другого (или несколько явлений имеют общие воздействующие факторы) [4].

Для сопоставления могут быть использованы как непараметрические, так и параметрические статистики. Первые применяются, например, для анализа рангов элементов сопоставляемых массивов, вторые – для анализа количественных значений элементов.

В данной работе используются перечни биграмм, имеющих наибольшее значение в соответствии со статической мерой T-score [5]. Выбор параметра сортировки обусловлен тем, что мера T-score позволяет выявить такие биграммы, которые характеризуют документ в целом, извлечь из текста наиболее частотные обороты – дискурсивные слова, предложно-падежные и глагольно-предложные конструкции, формулы, сложные предлоги, союзы, частицы. «Эта мера, – отмечают Е. В. Ягунова и Л. М. Пивоварова, – оказывается полезна при решении задачи о выделении тех единиц, которые характеризуют **все** (или **подавляющее большинство**) текстов коллекции. Основная масса таких сочетаний характеризует, скорее, особенности стиля текстов коллекции» [12] [6].

3.1. Коэффициент ранговой корреляции r-Спирмена

Одной из наиболее часто применяемых непараметрических статистик является статистика *r*-Спирмена, которая используется для сопоставления рангов:

$$r_s = 1 - \frac{6\sum_{i=1}^{n}\left(v_i - w_i\right)^2 + A + B}{N\left(N^2 - 1\right)},$$

где v_i – ранги элементов первого массива; w_i – ранги элементов второго массива; N – количество элементов в массиве; A, B – поправочные коэффициенты при наличии повторяющихся значений в массивах:

[3] См. раздел Источники в конце работы.

[4] Мерой корреляции служит коэффициент корреляции ρ(*r*), значение которого может находиться в интервале от –1 до +1. Значение 0 соответствует отсутствию статистической корреляционной связи. Значения +1, –1 соответствуют наличию сильной (функциональной) корреляции.

[5] $\text{T-score} = \dfrac{F\left(w_1, w_2\right) - \dfrac{F\left(w_1\right) \times F\left(w_2\right)}{N}}{\sqrt{F\left(w_1, w_2\right)}}$, где $F(w_1)$ – частота первого компонента; $F(w_2)$ – частота второго компонента; $F(w_1, w_2)$ – частота сочетания w_1w_2; N – общее число биграмм в корпусе. Мера считается наиболее эффективной для выделения статистически значимых биграмм, претендующих на статус устойчивых [13; 14; 15].

[6] Недостатком меры считается ее способность выявлять также и грамматически, и семантически несвязанные сочетания с наиболее частотными словами, например служебными [16]. В связи с отсутствием необходимости дифференцировать в сопоставляемых перечнях устойчивые и «случайные» сочетания считаем, что это свойство меры не является препятствием для использования наиболее частых биграмм, полученных на основе меры T-score, для корреляционного анализа.

$$A = \frac{n^3 - n}{12}, \quad B = \frac{k^3 - k}{12},$$

где n – число одинаковых рангов в первом массиве; k – число одинаковых рангов во втором массиве [17].

3.2. Коэффициент парной корреляции r-Пирсона

Для оценки близости биграмм на основе их статистического значения T-score может быть использована параметрическая статистика r-Пирсона:

$$r_{xy} = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum \cdot (y_i - \bar{y})^2}} = \frac{n\sum(x_i \cdot y_i) - \left(\sum x_i \cdot \sum y_i\right)}{\sqrt{\left[n \cdot \sum x_i^2 - \left(\sum x_i\right)^2\right] \cdot \left[n \cdot \sum y_i^2 - \left(\sum y_i\right)^2\right]}},$$

где x_i – значения, принимаемые переменной X; y_i – значения, принимаемые переменной Y; $\bar{x}$ – средняя по X; $\bar{y}$ – средняя по Y [18].

4. Материал для анализа: рукописи, подкорпуса, биграммы

Сопоставляемыми подкорпусами являются:

– коллекция полных апракосов, включающая три списка XII–XIII веков объемом 220 023 текстовые формы (лемматизированных форм – 129 135, 59 % от общего числа форм);

– коллекция кратких апракосов, включающая четыре списка XI–XIV веков объемом 164 429 текстовых форм (лемматизированных форм – 89 795, 55 % от общего числа форм);

– рукопись тетра XII века объемом 67 454 текстовые формы (лемматизированных форм – 32 064, 48 % от общего числа форм);

– рукопись Пантелеймонова Евангелия XII века объемом 68 734 текстовые формы (лемматизированных форм – 43 062, 63 % от общего числа форм).

Использование перечней биграмм для анализа близости подкорпусов с применением корреляционных методов основано на том, что одна и та же биграмма может иметь в сопоставляемых перечнях: а) различные ранги, б) разные значения T-score. Для анализа каждой биграмме перечня одного подкорпуса приписаны соответствующие ранги и значения T-score сопоставляемого перечня другого подкорпуса.

Анализу подвергнуты перечни биграмм с наибольшим значением T-score, компонентами которых являются леммы [7]; объем выборки – 300 элементов (первые 20 позиций перечней см. в прил. 1) [8]. Перечни биграмм, извлеченных из подкорпусов, сравниваются попарно (см. прил. 2).

Понятно, что состав перечней не идентичен, так как только часть биграмм совпадает в сопоставляемых перечнях из 300 элементов [9]. Поэтому проведено две серии экспериментов. В первой серии при отсутствии биграммы в сопоставляемом подкорпусе ей присвоены компенсирующие ранги [10] и значения [11] T-score (результаты, помеченные в разделе 5 литерой *а*). Во второй серии анализируются только такие биграммы, которые совпадают в сопоставляемых массивах (пункты, помеченные литерой *б*).

5. Результаты измерений [12]

5.1. Для сопоставления рангов биграмм использован коэффициент ранговой корреляции r-Спирмена.

5.1.1. Коэффициенты ранговой корреляции r-Спирмена для подкорпусов:

а) [13]

Подкорпуса	*N* [14]	*Rs*	*p*-value	Ранг
ПА & КА	300	0,592447	0,000000	1
ПА & Т	300	0,512577	0,000000	2
КА & Т	300	0,477041	0,000000	3

б) [15]

Подкорпуса	*N*	*Rs*	*p*-value	Ранг
ПА & КА	180	0,741679	0,000000	1
ПА & Т	175	0,731482	0,000000	2
КА & Т	170	0,661473	0,000000	3

[7] При выборе леммы как компонента биграммы мы руководствовались необходимостью максимально устранить графико-орфографическую вариативность лингвистических единиц. В то же время мы понимаем, что ограничение выборки только лемматизированными формами накладывает определенную степень вероятности на выводы.

[8] Выборка биграмм с наибольшим статистическим значением T-score осуществлена с помощью модуля n-грамм корпуса «Манускрипт» (URL: http://manuscripts.ru/mns/cred_ngr.stat); о модуле см., например, [19; 20].

[9] В ПА и КА совпадают 180 биграмм, в ПА и Т – 175, в КА и Т – 170; в ПЕ и ПА совпадают 190 биграмм, в ПЕ и КА – 178, в ПЕ и Т – 153 биграммы.

[10] Для ПА – 355, для КА – 361, для Т – 373, выбранные в соответствии с количеством несовпадающих биграмм: *количество_биграмм_в_перечне* + *количество_несовпадающих_биграмм*/2; например, для ПА: 300 + 110/2 = 355.

[11] Для ПА – 3,202, для КА – 2,609, для Т – 1,651, полученные экспоненциальной аппроксимацией значений биграмм в каждом из подкорпусов.

[12] Расчеты были осуществлены с помощью программы Statistica (StatSoft Russia).

[13] Результаты анализа данных с компенсирующими значениями для биграмм, отсутствующих в сопоставляемом подкорпусе.

[14] Количество сопоставляемых биграмм.

[15] Результаты анализа биграмм, зафиксированных в обоих сопоставляемых массивах данных.

Оба измерения выявили наибольшую близость между подкорпусами ПА и КА, наименьшую – между КА и Т (см. столбец *Ранг*).

5.1.2. Коэффициент ранговой корреляции *r*-Спирмена для ПЕ и подкорпусов:

а)

Подкорпуса	*N*	*Rs*	*p*-value	Ранг
ПЕ & ПА	300	0,636189	0,000000	1
ПЕ & КА	300	0,626844	0,000000	2
ПЕ & Т	300	0,436115	0,000000	3

б)

Подкорпуса	*N*	*Rs*	*p*-value	Ранг
ПЕ & ПА	190	0,738369	0,000000	1
ПЕ & КА	178	0,697511	0,000000	3
ПЕ & Т	153	0,717090	0,000000	2

В случае измерения 300 биграмм обе статистики демонстрируют наибольшую близость друг другу ПЕ и подкорпуса ПА, наименьшую – ПЕ и Т.

В случае измерения только совпадающих биграмм обе статистики также демонстрируют наибольшую близость друг другу ПЕ и подкорпуса ПА, но наименьшую – ПЕ и КА (см. столбец *Ранг*).

Во всех случаях статистические значения имеют максимально высокую достоверность (значение p-value значительно меньше 0,05).

5.2. Для сопоставления значений, полученных с помощью меры T-score, использована статистика *r*-Пирсона.

5.2.1. Коэффициент парной корреляции *r*-Пирсона для подкорпусов:

а)

Подкорпуса	*N*	*Rs*	*p*-value	Ранг
ПА & КА	300	0,723550	0,000000	1
ПА & Т	300	0,649952	0,000000	2
КА & Т	300	0,620782	0,000000	3

б)

Подкорпуса	*N*	*Rp*	*p*-value	Ранг
ПА & КА	180	0,638414	0,000000	3
ПА & Т	175	0,841860	0,000000	1
КА & Т	170	0,821947	0,000000	2

Результаты измерений существенно различны: при использовании компенсирующих значений наибольшую близость обнаруживают ПА и КА, наименьшую – КА и Т; при измерении только совпадающих биграмм наибольшая близость выявляется между ПА и Т, наименьшая – между ПА и КА (см. столбец *Ранг*).

5.2.2. Коэффициент парной корреляции *r*-Пирсона для ПЕ и подкорпусов:

а)

Корпуса	*N*	*Rp*	*p*-value	Ранг
ПЕ & ПА	300	0,796651	0,00	1
ПЕ & КА	300	0,784099	0,00	2
ПЕ & Т	300	0,612462	0,00	3

б)

Корпуса	*N*	*Rp*	*p*-value	Ранг
ПЕ & ПА	190	0,740109	0,00	2
ПЕ & КА	178	0,558958	0,00	3
ПЕ & Т	153	0,879215	0,00	1

И в этом случае статистические коэффициенты демонстрируют различную близость ПЕ и подкорусов: при использовании компенсирующих значений наибольшая близость выявляется между ПЕ и ПА, наименьшая – между ПЕ и Т; сравнение только совпадающих биграмм дает наибольшую близость между ПЕ и Т и наименьшую – между ПЕ и КА (см. столбец *Ранг*).

6. Обсуждение результатов измерений

Результаты оценки близости подкорпусов между собой, с одной стороны, и ПЕ и подкорпусов, с другой, с помощью непараметрической ранговой статистики *r*-Спирмена и параметрической статистики *r*-Пирсона как достаточно близки, так существенно различны.

6.1. Близость оценок проявляется в том, что все коэффициенты указывают на наличие корреляционной связи между сопоставляемыми массивами биграмм, а также на достаточно существенную степень их корреляции – от 0,43 (п. 5.1.2а), что соответствует умеренной корреляции по таблице Чеддока, до 0,88 (п. 5.2.2б), что соответствует высокой силе связи; большая часть значений лежит в диапазоне от 0,5 до 0,7, что характеризует тесноту связи как заметную.

Понятно, что и наличие корреляционной связи, и ее сила связаны с тем, что подкорпуса и сопоставляемые тексты являются Евангелиями.

6.2. Оценка силы связи между подкорпусами позволяет говорить о различиях в степени близости сопоставляемых массивов биграмм.

6.2.1. Значения коэффициентов *r*-Спирмена для подкорпусов позволяют сделать вывод о большей близости ПА и КА и наименьшей КА и Т, это обнаруживается как при сопоставлении массивов из 300 биграмм, так и при сравнении массивов, имеющих только идентичные биграммы (п. 5.1.1а, б).

Значения коэффициентов ранговой корреляции ПЕ и подкорпусов показывают наибольшую близость ПЕ и ПА как при сопостав-

лении 300 биграмм, так и при сопоставлении совпадающих биграмм (п. 5.1.2а, б – ранг 1), но одновременно в первом случае наименьшая близость обнаруживается между ПЕ и Т, во втором – между ПЕ и КА (п. 5.1.2а, б – ранг 3).

6.2.2. Сопоставление отношений, выявленных на основе ранговой корреляции, между подкорпусами Евангелий разного типа, с одной стороны, и между ПЕ и подкорпусами, с другой, позволяет увидеть корреляцию между максимальной близостью ПА & КА и ПЕ & ПА, что может быть объяснено тем, что ПА является полным апракосом.

Одновременно с этим отношения ПЕ & КА и ПЕ & Т, оцениваемые не так однозначно (ПЕ имеет наименьшую близость или с Т, или с КА), как будто соотносятся с наименьшей близостью КА и Т.

6.2.3. Коэффициенты статистики *r*-Пирсона как при измерении отношений между подкорпусами рукописей разного типа, так и между ПЕ и подкорпусами выше, чем в ранговой статистике *r*-Спирмена (пп. 5.2.1 и 5.2.2).

Оценка отношений между подкорпусами на основе значений биграмм при использовании компенсирующих значений T-score (перечень из 300 биграмм) совпадает с оценкой на основе рангов (п. 5.1.1а и 5.2.1а). Точно так же совпадают результаты оценки отношений между ПЕ и подкорпусами (п. 5.1.2а и 5.2.2а).

6.2.4. Иные результаты дает статистика *r*-Пирсона при использовании только тех биграмм, которые совпадают в сопоставляемых массивах. Наибольшая близость выявлена между ПА и Т, наименьшая – между ПА и КА (п. 5.2.1б), что соотносится с результатами оценки отношений между ПЕ и подкорпусами: наибольшая близость между ПЕ и Т, наименьшая – между ПЕ и КА (п. 5.2.2б).

7. Выводы

Понятно, что результаты экспериментов всегда зависят от условий их проведения. В нашем случае на результат, безусловно, повлияли неполная лемматизация текстов и несовпадение перечней биграмм в сопоставляемых подкорпусах.

Тем не менее доказано, что между рядами биграмм подкорпусов существует статистическая корреляция, имеющая высокую статистическую вероятность.

Считаем, что выявленная идентичная оценка отношений между подкорпусами, например между ПА и КА (пп. 5.1.1а, б, 5.2.1а), между ПЕ и ПА (пп. 5.1.2а, б, 5.2.2а), позволяет сделать вывод, что неполнота данных не была столь существенна, как это можно было бы предполагать.

В то же время, если принять во внимание, что параметрические статистики обладают большей силой по сравнению с ранговыми и что перечни без компенсирующих рангов и значений должны были дать более объективные результаты, несмотря на меньший объем, то следует с большей долей доверия отнестись к выводам п. 6.2.4.

Следует признать, что причины обнаруженных расхождений в оценке одних и тех же отношений нужно искать не только в неизбежной погрешности измерений, но и в других факторах: и в первую очередь – в принадлежности Евангелий одного типа к различным редакциям, во временных и локальных их характеристиках[16]. И эти факторы могут быть обнаружены, в частности, с помощью применения корреляционного анализа к перечням биграмм каждой пары рукописей. При этом полученные в данной работе корреляционные коэффициенты подкорпусов евангельских списков разных типов могут использоваться как нормирующие.

Источники

Полные апракосы

Евангелие апракос полный («Пантелеймоново Евангелие»), РНБ, Соф. 1, XII–XIII в., 224 л.

Евангелие апракос полный («Мстиславово Евангелие»), ГИМ, Син. 1203, до 1117 г., 213 л.

Евангелие апракос полный, РГБ, Рум. 104, кон. XII – нач. XIII (?) в., 158 л.

Евангелие апракос полный («Симоновское евангелие»), РГБ, Рум. 105, 1270 г., 167 л.

Краткие апракосы

Евангелие апракос краткий («Остромирово Евангелие»), РНБ, F.п.1.5., 1056–1057 гг., 294 л.

Евангелие апракос краткий («Архангельское евангелие»), РГБ, М.1666, 1092 г., 178 л.

Евангелие апракос краткий (Погодинское Евангелие), РНБ, Погод. 11, XI|XII|XII–XIII в., 264 л.

Евангелие апракос краткий, РНБ, F.n.I.13, кон. XIII – нач. XIV в., 75 л.

Евангелие тетр

Евангелие тетр («Типографское евангелие»), РГАДА, ф. 381 (Син. тип.), № 1, XII в., 193 л.

Библиографические ссылки

1. *Воскресенский Г. А.* Характеристические черты четырех редакций славянского перевода евангелия от

[16] Например, неоднократно было показано, что в одну и ту же группу, выделенную на основании анализа текстологических различий или лингвистических разночтений, входят как полные и краткие служебные Евангелия, так и тетры [21; 22; 23; 24], а древнейшие русские евангельские списки одного типа не представляют единства ни в текстологи-ческом, ни в лингвистическом отношении [25].

Марка по сто двенадцати рукописям евангелия XI–XVI вв. М., 1896. 304 с.

2. *Жуковская Л. П.* Текстология и язык древнейших славянских памятников. М. : Наука, 1976. 368 с.

3. *Алексеев А. А.* Текстология славянской Библии. СПб. : Дмитрий Буланин, 1999. 256 с.

4. *Алексеев А. А., Кузнецова Е. Л.* ЭВМ и проблемы текстологии древнеславянских текстов // Лингвистические задачи и обработка данных на ЭВМ. М. : ИРЯ АН СССР, 1987. С. 111–121.

5. *Баранов В. А.* N-граммы Пантелеймонова Евангелия (РНБ, Соф. 1) на фоне древнерусских евангельских списков // Лингвокультурологические исследования развития русского языка в условиях полиэтнической среды: опыт и перспективы : тр. и матер. : в 2 т. / под общ. ред. Е. А. Горобец, О. Ф. Жолобова, М. О. Новак. Казань, Изд-во Казан. ун-та, 2018. Т. 2. С. 21–25.

6. *Азарова И. В., Алексеева Е. Л., Миронова Д. М.* Кластеризация рукописей на базе совпадения разночтений как основа публикации славянской традиции // Материалы XLIII междунар. филол. конф. : секция прикладной и математической лингвистики, 11–15 марта 2014 г. / [отв. ред. М. В. Хохлова]. СПб. : Филол. фак. СПбГУ, 2014. С. 10–22.

7. *Миронова Д. М.* Автоматизированная классификация древних рукописей (На материале 525 списков славянского Евангелия от Матфея XI–XVI вв.) : дис. … канд. филол. наук : спец. 10.02.21. – Прикладная и математическая лингвистика. СПб., 2018. 315 с.

8. *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.

9. *Жуковская Л. П.* Указ. соч. С. 224–349.

10. *Горина Н. Л.* Опыт оценки текстологической значимости разночтений // Труды Отдела древнерусской литературы / Российская академия наук. Институт русской литературы (Пушкинский Дом); Гл. ред. серии Д. С. Лихачев, Ред.: А. А. Алексеев, М. А. Салмина. СПб. : Дмитрий Буланин, 1996. Т. 49. С. 223–338.

11. *Миронова Д. М.* Указ. соч.

12. *Ягунова Е. В., Пивоварова Л. М.* Указ. соч.

13. *Evert, S.* Association Measures. Computational Approaches to Collocations. Section 4.3. Available at: http://collocations.de/AM/index.html (accessed 15.10.2018).

14. *Кочеткова Н. А.* Статистические языковые методы. Коллокации и коллигации // Новые информационные технологии в автоматизированных системах. 2013. № 16 [Электронный ресурс]. С. 301–305. URL: http://cyberleninka.ru/article/n/statisticheskie-yazykovye-metodykollokatsii-i-kolligatsii (дата обращения: 10.10.2018).

15. *Ягунова Е. В., Пивоварова Л. М.* От коллокаций к конструкциям // Русский язык: конструкционные и лексико-семантические подходы / Отв. ред. С. С. Сай. СПб., 2013. (Acta Linguistica petropolitana: Труды Института лингвистических исследований РАН). URL: https://bit.ly/2OWkAmC (дата обращения: 06.10.2018).

16. *Хохлова М. В.* Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia 34. Инструментарий русистики: Корпусные подходы / под ред. А. Мустайоки, М. В. Копотева, Л. А. Бирюлина, Е. Ю. Протасова. Хельсинки, 2008. С. 343–357. С. 348.

17. *Фёрстер Э., Рёнц Б.* Методы корреляционного и регрессионного анализа. Руководство для экономистов / пер. с нем. и предисл. В. М. Ивановой. М. : Финансы и статистика, 1983. 304 с. С. 160–163.

18. Там же.

19. *Баранов В. А.* Модуль n-грамм исторического корпуса «Манускрипт»: структурные и лингвистические параметры // Научное наследие В. А. Богородицкого и современный вектор исследований Казанской лингвистической школы : тр. и матер. междунар. конф. (Казань, 31 окт. – 3 нояб. 2016 г.) : в 2 т. / под общ. ред. К. Р. Галиуллина, Е. А. Горобец, Г. А. Николаева. Казань: Изд-во Казан. ун-та, 2016. Т. 1. С. 50–61.

20. *Баранов В. А.* Количественный и статистический анализ средневековых славянских текстов: инструментарий корпуса «Манускрипт» и методика его использования // Цифровая гуманитаристика: ресурсы, методы, исследования: материалы Междунар. науч. конф. (г. Пермь, 16–18 мая 2017 г.): в 2 ч. / Перм. гос. нац. исслед. ун-т. Пермь, 2017. Ч. 1. С. 40–49.

21. *Воскресенский Г. А.* Указ. соч. С. 12–24, 31–46.

22. *Жуковская Л. П.* Указ. соч. С. 332.

23. *Горина Н. Л.* Указ. соч. С. 327.

24. *Миронова Д. М.* Указ. соч. С. 183–189.

25. *Жуковская Л. П.* Указ. соч. С. 259, 348 и др.

References

1. Voskresenskii G.A. *Kharakteristicheskie cherty chetyrekh redaktsii slavyanskogo perevoda evangeliya ot Marka po sto dvenadtsati rukopisyam evangeliya XI–XVI vv.* [Characteristic features of the four editions of the Slavic Gospel translation from Mark on one hundred and twelve Gospel manuscripts of the 11th – 16th centuries]. Moscow, 1896, 304 p. (in Russ.).

2. Zhukovskaya L.P. *Tekstologiya i yazyk drevneishikh slavyanskikh pamyatnikov* [Textology and language of the oldest Slavic manuscripts]. Moscow : Nauka, 1976, 368 p. (in Russ.).

3. Alekseev A.A. *Tekstologiya slavyanskoi Biblii* [Textology of the Slavic Bible]. St. Petersburg, Dmitry Bulanin, 1999, 256 p. (in Russ.).

4. Alekseev A.A., Kuznetsova E.L. *EVM i problemy tekstologii drevneslavyanskikh tekstov* [Computer and the problems of textual Old Slavonic texts]. *Lingvisticheskie zadachi i obrabotka dannykh na EVM* [Linguistic tasks and data processing on a computer]. Moscow, 1987, pp. 111–121 (in Russ.).

5. Baranov V.A. *N-grammy Panteleimonova Evangeliya (RNB, Sof. 1) na fone drevnerusskikh evangel'skikh spiskov* [The N-grams of the Panteleimon Gospel (NNB, Sof. 1) against the backdrop of ancient Russian evangelical manuscripts]. *Lingvokul'turologicheskie issledovaniya razvitiya russkogo yazyka v usloviyakh*

polietnicheskoi sredy : opyt i perspektivy : trudy i materialy : v 2 t. [Linguocultural studies of the development of the Russian language in a multi-ethnic environment: experience and prospects : works and materials : in 2 volumes]. (eds. E. A. Gorobets, O. F. Zholobova, M. O. Novak). Kazan', 2018, vol. 2, pp. 21–25 (in Russ.).

6. Azarova I.V., Alekseeva E.L., Mironova D.M. *Klasterizatsiya rukopisei na baze sovpadeniya raznochtenii kak osnova publikatsii slavyanskoi traditsii* [Clustering of manuscripts based on the coincidence of different readings as the basis for the publication of the Slavic tradition]. *Materialy XLIII mezhdunar. filol. konf. : sektsiya prikladnoi i matematicheskoi lingvistiki, 11–15 marta 2014 g.* [Proceedings of the XLIII International Philological Conference : Section of Applied and Mathematical Linguistics, March 11–15, 2014] (ed. M. V. Khokhlova). St. Petersburg, 2014, pp. 10–22 (in Russ.).

7. Mironova D.M. *Avtomatizirovannaya klassifikatsiya drevnikh rukopisei (Na materiale 525 spiskov slavyanskogo Evangeliya ot Matfeya XI–XVI vv.)* [Automated classification of ancient manuscripts (On the material of 525 manuscripts of the Slavic Gospel of Matthew XI–XVI centuries.)]. PhD thesis. St. Petersburg, 2018, 315 p. (in Russ.).

8. Baranov V.A. *Istoricheskii korpus kak tsel' i instrument korpusnoi paleoslavistiki* [Baranov V. A. The Historical Corpus as a Purpose and Instrument of Corps Paleoslavistics]. Scripta & e-Scripta : The Journal of Interdisciplinary Mediaeval Studies. Vol. 14–15. Sofia : "Boyan Penev" Publishing Center; Institute of Literature, BAS, 2015, pp. 39–62 (in Russ.).

9. Zhukovskaya L.P. *Opere Citato.* Pp. 224–349.

10. Gorina N.L. *Opyt otsenki tekstologicheskoi znachimosti raznochtenii* [Experience in assessing the textual significance of discrepancies]. *Trudy Otdela drevnerusskoi literatury / Rossiiskaya akademiya nauk. Institut russkoi literatury (Pushkinskii Dom)* [Proceedings of the Department of Old Russian Literature / Russian Academy of Sciences. Institute of Russian Literature (Pushkin House)]. (Ch. ed. D. Likhachev, eds. A. A. Alekseev, M. A. Salmina). Vol. 49. St. Petersburg : Dmitry Bulanin, 1996, pp. 223–338 (in Russ.).

11. Mironova D.M. *Opere Citato.*

12. Yagunova E.V., Pivovarova L.M. *Opere Citato.*

13. Evert S. Association Measures. Computational Approaches to Collocations. Section 4.3. Available at: http://collocations.de/AM/index.html (accessed 15.10.2018).

14. Kochetkova N.A. *Statisticheskie yazykovye metody. Kollokatsii i kolligatsii* [Statistical language methods. Collocations and Colligations]. Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh [New information technologies in automated systems.]. 2013, № 16, pp. 301–305 (in Russ.). Available at: http://cyberleninka.ru/article/n/statisticheskie-yazykovye-metodykollokatsii-i-kolligatsii (accessed 10.10.2018).

15. Yagunova E.V., Pivovarova L.M. Ot kollokatsii k konstruktsiyam [From collocations to syntax]. Russkii yazyk: konstruktsionnye i leksiko-semanticheskie podkhody [Russian language: constructional and lexical-semantic approaches] (ed. S. S. Sai). St. Petersburg, 2013. (Acta Linguistica petropolitana: Trudy Instituta lingvisticheskikh issledovanii RAN [Proceedings of the Institute of Linguistic Studies of the Russian Academy of Sciences]) (in Russ.). Available at: https://bit.ly/2OWkAmC (accessed 06.10.2018).

16. Khokhlova M.V. *Eksperimental'naya proverka metodov vydeleniya kollokatsii* [Experimental verification of collocation methods]. Slavica Helsingiensia 34. Instrumentarii rusistiki: Korpusnye podkhody [Tools of Russian Studies: Corpus Approaches] (eds. A. Mustaioki, M.V. Kopoteva, L.A. Biryulina, E.Yu. Protasova). Khel'sinki, 2008, pp. 343–357, p. 348 (in Russ.).

17. Ferster E., Rents B. *Metody korrelyatsionnogo i regressionnogo analiza. Rukovodstvo dlya ekonomistov* [Methods of Correlation and Regression Analysis. Manual for Economists]. Moscow, 1983, 304 p., pp. 160–163.

18. Ibid.

19. Baranov V.A. *Modul' n-gramm istoricheskogo korpusa «Manuskript»: strukturnye i lingvisticheskie parametry* [The n-grams module of the historic Manuscript corpus: structural and linguistic parameters]. *Nauchnoe nasledie V. A. Bogoroditskogo i sovremennyi vektor issledovanii Kazanskoi lingvisticheskoi shkoly : truda i materialy mezhdunarodnoi konferentsii (Kazan', 31 okt. – 3 noyab. 2016 g.): v 2 t.* [The scientific heritage of V. A. Bogoroditsky and the modern vector of research of the Kazan linguistic school : proceedings and materials of the international conference (Kazan, October 31 – November 3, 2016) : in 2 volumes]. (eds. K. R. Galiullina, E. A. Gorobets, G. A. Nikolaeva). Vol. 1. Kazan, 2016, pp. 50–61.

20. Baranov V.A. *Kolichestvennyi i statisticheskii analiz srednevekovykh slavyanskikh tekstov: instrumentarii korpusa «Manuskript» i metodika ego ispol'zovaniya* [Quantitative and statistical analysis of medieval Slavic texts: tools of the Manuscript corpus and the method of its use]. *Tsifrovaya gumanitaristika: resursy, metody, issledovaniya : materialy Mezhdunarodnoi nauchnoi konferentsii (g. Perm', 16–18 maya 2017 g.) : v 2 ch.* [Digital humanities: resources, methods, research : materials of the International Scientific Conference (Perm, May 16–18, 2017) : at 2 vol.]. Vol. 1. Perm, 2017, pp. 40–49.

21. Voskresenskii G.A. *Opere Citato.* Pp. 12–24, 31–46.

22. Zhukovskaya L.P. *Opere Citato.* P. 332.

23. Gorina N.L. *Opere Citato.* P. 327.

24. Mironova D.M. *Opere Citato.* Pp. 183–189.

25. Zhukovskaya L.P. *Opere Citato.* Pp. 259, 348, etc.

Приложение 1. **Перечни первых 20 биграмм, имеющих наибольшее значение в соответствии с мерой T-score**

	Пантелеймоново Евангелие			*Полные апракосы*			*Краткие апракосы*			*Тетр*		
Ранг	Биграмма	F	T-score	Биграмма	F	T-score	Биграмма	F	T-score	Биграмма	F	T-score
1	онъ же	223	13,959	онъ же	694	24,249	онъ же	428	18,991	онъ же	199	13,282
2	глаголати въı	151	11,549	ннкъто же	283	16,334	же рєщи	259	12,883	оуученикъ свои	66	7,907
3	рєщи господь	138	11,349	оно врема	247	15,624	отъ лоука	167	12,463	отъ лоука	61	7,525
4	ннкъто же	91	9,282	ннъто же	232	14,796	отъ иоанъ	163	12,348	же рєщи	82	7,417
5	же рєщи	123	8,838	оуученикъ свои	228	14,593	врема оно	136	11,495	въ домъ	67	7,324
6	оуученикъ свои	72	8,187	отъ иоанъ	176	12,725	глаголати въı	149	11,193	отъ маркъ	52	7,083
7	ннъто же	68	8,028	глаголати въı	191	12,455	въ врема	152	11,105	на небо	51	6,928
8	въ домъ	69	7,303	отъ лоука	162	12,072	рєщи господь	133	11,014	иднъ отъ	55	6,869
9	отьць мои	56	7,286	отьць мои	152	12,020	въ оно	243	10,889	не мощи	51	6,866
10	амнь глаголати	48	6,783	иднъ отъ	155	11,125	отьць мои	125	10,877	ннкъто же	47	6,687
11	исоусъ же	70	6,495	вьсь мнръ	133	11,022	оуученикъ свои	106	9,847	на земла	43	6,224
12	вьсь мнръ	46	6,492	исоусъ же	188	10,969	же родити	105	9,711	не вѣдѣти	44	6,189
13	на небо	42	6,371	отъ марко	119	10,654	въ домъ	118	9,652	ннъто же	39	6,097
14	на земла	43	6,362	же рєщи	175	10,184	посълати азъ	100	9,477	съ ради	38	5,992
15	аще къто	42	6,172	свои оуученикъ	116	10,060	азъ бъıти	206	9,265	азъ бъıти	82	5,910
16	же родити	40	6,057	посълати азъ	114	10,048	рєщи же	170	9,076	иже бъıти	71	5,862
17	посълати азъ	41	6,023	азъ бъıти	248	9,812	иднъ отъ	101	8,983	же родити	37	5,830
18	въ нма	49	6,009	въı яко	161	9,800	рєщи къ	108	8,950	не имѣти	45	5,826
19	иже бъıти	106	5,924	съ ради	104	9,672	иже бъıти	173	8,817	аще къто	36	5,700
20	прити къ	39	5,888	не мощи	107	9,634	прити къ	86	8,800	посълати азъ	35	5,593

Приложение 2. **Перечень первых 20 биграмм, их ранги и статистические значения (выровнено по ПЕ)**

№	N-грамма	Ранг ПЕ	Ранг ПА	Ранг КА	Ранг Т	T-score ПЕ	T-score ПА	T-score КА	T-score Т
1	онъ же	1	1	1	1	13,959	24,249	18,991	13,282
2	глаголати бы	2	7	6	373	11,549	12,455	11,193	2,310
3	рѣшн господь	3	38	8	373	11,349	8,329	11,014	2,310
4	ннкъто же	4	2	99	10	9,282	16,334	5,244	6,687
5	же рѣшн	5	14	2	4	8,838	10,184	12,883	7,417
6	оучєнникъ свон	6	5	11	2	8,187	14,593	9,847	7,907
7	ннчьто же	7	4	115	13	8,028	14,796	4,977	6,097
8	въ домъ	8	25	13	5	7,303	9,071	9,652	7,324
9	отьць мон	9	9	10	21	7,286	12,020	10,877	5,559
10	амнньъ глаголати	10	41	30	373	6,783	8,137	7,643	2,310
11	нсоусъ же	11	12	84	373	6,495	10,969	5,576	2,310
12	вьсь мнръ	12	11	37	59	6,492	11,022	7,407	4,060
13	на небо	13	282	56	373	6,371	4,419	6,338	2,310
14	на зємлѧ	14	55	44	11	6,362	7,257	6,750	6,224
15	аще къто	15	50	23	19	6,172	7,719	8,493	5,700
16	же родити	16	44	12	17	6,057	8,077	9,711	5,830
17	посълати азъ	17	16	14	20	6,023	10,048	9,477	5,593
18	въ нмѧ	18	40	34	22	6,009	8,180	7,451	5,472
19	нже бъіти	19	49	19	16	5,924	7,734	8,817	5,862
20	прнити къ	20	36	20	33	5,888	8,400	8,800	4,724

* * *

Correlation analysis of the bigrams of the copies of Russian Gospels of the 11th – 14th Centuries

V. A. Baranov, DSc in Philology, Professor, Kalashnikov ISTU, Izhevsk, Russia

With the aid of the correlation analysis the paper considers the degree of closeness of the copies of Russian Gospels of the 11th – 14th century of different types (complete aprakoses, short aprakoses, Four Gospels) to one another, and also the degree of closeness of the Panteleymon Gospel of the 12th century (complete aprakos) to each of the types. The lists of bigrams with the greatest value of T-score whose components are lemmas has been analyzed; the amount of retrievals from each subcorpus has been 300 elements. To reveal the closeness of the lists, the nonparametric statistic r-Spearman was applied to the ranks of bigrams and the parametric statistic r-Pearson to their T-score values.

The obtained results indicate the presence of the correlation relationship between the compared arrays of bigrams which has a high statistic probability and also a sufficiently significant degree of their correlation which is characterized either like moderate or noticeable. The assessment of the relationship force between the subcorpora suggests differences in the degree of closeness of the compared arrays of bigrams. In accordance with the rank correlation r-Spearman, the highest closeness is shown by the subcorpus of complete aprakoses and the subcorpus of short aprakoses, the Panteleymon Gospel and complete aprakoses, the lowest by short aprakoses, Four Gospels, the Panteleymon Gospel and Four Gospels (or short aprakoses). In accordance with the correlation r-Pearson the highest closeness is revealed between the complete aprakoses and Four Gospels, the lowest between the complete and short aprakoses. The relationships of the Panteleymon Gospel with Four Gospels and short aparakoses are similar to the relationships of the complete aprakoses with them.

Keywords: linguistic statistics, bigram, Gospel, Russian manuscripts, Gospel Book, aprakos evangelium, Panteleymon Gospel.

Получено: 22.10.18