УДК 519.2:801.82(045)
DOI: 10.22213/2410-9304-2019-4-63-77

ДРЕВНЕРУССКИЕ РУКОПИСИ КАК ОБЪЕКТ СТАТИСТИЧЕСКОГО АНАЛИЗА*

В. А. Баранов, доктор филологических наук, профессор, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия
О. Ф. Жолобов, доктор филологических наук, профессор, Казанский федеральный университет, Казань, Россия

В работе описаны два статистических эксперимента, целью которых стало выявление корреляционной близости / удаленности 12 текстов, дошедших до нас в русских списках XI века, и сопоставление с ними произведений автора XII века Кирилла Туровского (РНБ, F.п.I. 39, XIII в.; лл. 1–48), приведены результаты сопоставительного анализа: а) различных способов извлечения лингвистических единиц из текстов и б) выборок разного объема, а также лингвистической интерпретации основных закономерностей группировки рукописей.

Степень лингвостатистической тесноты рукописей вычисляется в два этапа: на первом сопоставляются перечни наиболее частотных слов каждой пары текстов (вычисляется коэффициент ранговой корреляции Спирмена), на втором тексты группируются на основе полученных значений корреляции, которые принимаются за расстояния между рукописями (используется кластерный анализ и строится дендрограмма).

Извлечение наиболее частотных слов рукописей, построение ранжированных перечней, получение сведений о количестве (а соответственно, о ранге) каждой из форм в других кодексах выполнено с помощью модуля статистики исторического корпуса «Манускрипт». Вычисление коэффициентов корреляции текстов и кластеризация текстов осуществлены с помощью программного пакета «Статистика» (TIBCO Software Inc.). Проанализированы перечни разного объема (от 50 до 300 словоформ), состоящие из единиц разной степени унификации относительно текстовых форм.

Результатом первого эксперимента стало выявление трех основных устойчивых кластеров подкорпуса – группы Евангелий, группы миней и группы сборников разного содержания.

Второй эксперимент дал возможность увидеть зависимость близости проповедей Кирилла Туровского разным кластерам от степени унификации форм в выборках и объема последних.

Лингвистический анализ результатов позволил выявить лексико-грамматические и лексико-семантические факторы, определяющие вхождение текстов Кирилла Туровского при различных исходных условиях выборки в разные кластеры – в группу Евангельских списков (при объеме выборки 50 или 100 слов), в подгруппу сборников (при выборке в 200 слов), в подгруппу Изборника 1073 г. и Пандектов Антиоха (выборка – 300 слов).

Ключевые слова: лингвистическая статистика, древнерусские тексты XI века, Кирилл Туровский.

Статистические методы и лингвистика

Применение количественных и статистических методов в лингвистических исследованиях имеет давнюю традицию, восходящую к 50-м годам прошлого века.

Важным событием в русистике стало появление в начале 70-х годов монографии Б. Н. Головина «Язык и статистика», целью которой стал анализ грамматической и стилистической системы русского литературного языка, отраженной в текстах XIX и XX веков [1]. И уже тогда, около 50 лет назад, автор, формулируя перспективы использования статистических методов в лингвистике, говорит о возможности их применения и для решения вопросов истории русского языка [2]. А в 1999 году под редакцией Л. И. Бородкина и И. М. Гарсковой выходит коллективная монография «Компьютеризованный статистический анализ для историков», в которой демонстрируется эффективность и результативность анализа российских и советских исторических письменных источников с помощью статистических методов (см., например, коллективную работу [3]).

В настоящее время использование статистических приемов при работе с современными текстами – практически обязательное условие теоретических и прикладных лингвистических работ. В то же время имеющееся значительное отставание в применении таких методов по отношению к средневековым славянским письменным памятникам объясняется отсутствием до недавнего времени массового текстового материала в машиночитаемой форме.

Сегодня в интернете доступно несколько специализированных ресурсов, содержащих транскрипции текстов: исторические коллекции Национального корпуса русского языка[1], корпус древнерусских берестяных грамот[2], Регенсбургский русский диахронический корпус[3], корпус нескольких списков Повести временных лет, подготовленный Дейвидом Бирнбаумом[4], подкорпуса древнерусских текстов университета в Тромсё[5] и проекта

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (РФФИ) в рамках проектов «Лингвостатистический анализ однокомпонентных и многокомпонентных лексических единиц исторического корпуса “Манускрипт”» (проект № 18-012-00463) (статистический анализ) и «Подготовка интернет-издания и комплексное исследование языка и письма Толстовского сборника XIII в. (РНБ, F.п.I.39)» (проект № 18-012-00428) (лингвистический анализ).

[1] Церковнославянский корпус. URL: http://www.ruscorpora.ru/search-orthlib.html; Древнерусский корпус. URL: http://www.ruscorpora.ru/search-old_rus.html; Старорусский корпус. URL: http://www.ruscorpora.ru/search-mid_rus.html; Корпус берестяных грамот. URL: http://www.ruscorpora.ru/search-birchbark.html.

[2] Корпус древнерусских берестяных грамот. URL: http://gramoty.ru/.

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

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

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

TITUS[6], корпус агиографических текстов СКАТ[7] и два корпуса проекта «Манускрипт: славянское письменное наследие» – исторический корпус «Манускрипт»[8] и корпус языка М. В. Ломоносова[9]. Эти ресурсы, созданные на основе разного текстового материала, имеющие различные поисковые возможности и направленные на решение разных задач, имеют идентичную направленность – исследовательскую, аналитическую.

Корпус «Манускрипт». Его модуль статистики и электронная коллекция русских рукописей XI века

Корпус «Манускрипт» (manuscripts.ru) содержит размеченные транскрипции средневековых славянских рукописей X–XV веков, снабжен формами запросов и вывода данных различного назначения (см. о корпусе, например, [4; 5; 6]). Одним из специализированных инструментов подготовки выборок является модуль статистики[10], позволяющий получить количественную и статистическую информацию о лингвистических единицах – текстовых формах и леммах [7], сравниваемых или сопоставляемых с контрастным подкорпусов[11].

Корпус имеет в своем составе коллекцию древнейших русских кодексов и отрывков XI века. Состав подкорпуса: два Евангелия (1056–1057 г., 1092 г.), Псалтырь, четыре служебные минеи на разные месяцы (три из которых – 1095–1096, 1096, 1097 гг. – из одного комплекта), два сборника с текстами разного содержания различных авторов (1073 и 1076 гг.), переводы сочинений Григория Богослова (IV в.), Антиоха Черноризца (VII в.) и Иоанна Мосха (VII в.) [8: №№ 3, 4, 5, 6, 7, 8, 9, 21, 24, 26, 31, 33][12] (см. Источники).

Корпус «Манускрипт» содержит также и другие древнейшие рукописи и тексты, несколько из которых – проповеди выдающегося русского писателя XII века Кирилла Туровского, включенные в сборник XIII века, – также использованы в работе. Тексты проповедей уже были объектом статистического анализа. В работе [9] показано, что наиболее близкими друг другу с точки зрения статистической значимости 15 наиболее частотных форм являются оригинальные тексты проповедей Кирилла Туровского, переводные тексты наставлений Ефрема Сирина и списков Апостола, а также частично Евангелий, Псалтыри и Летописей.

Цели, задачи, методы и материал

Цели работы – выявить статистическую (корреляционную) близость / удаленность текстов, дошедших до нас в русских списках XI века, и сопоставить с этими текстами произведения Кирилла Туровского, автора XII века. *Задачи* – установить влияние на результаты анализа: а) применения различных способов извлечения лингвистических единиц из текстов, б) объема выборок, в) дать лингвистическую интерпретацию основным статистическим закономерностям группировки рукописей.

Для сопоставления текстов в работе применены корреляционный и кластерный статистические методы.

Материалом для анализа послужили текстовые формы различной степени унификации русских списков XI века (проанализированы все 12 полных кодексов из сохранившихся от XI века рукописей, созданных на Руси) и проповедей Кирилла Туровского по наиболее близкому к оригиналу списку XIII века (РНБ, F.п.I. 39; лл. 1–48) [10].

Корреляционный анализ

Как известно, целью корреляционного анализа является обнаружение степени зависимости (тесноты) между двумя или более переменными, при которой изменения значений одной из них сопровождается изменением значений другой. Мерой корреляции является коэффициент корреляции r, значение которого изменяется от –1 при отрицательной корреляции до +1 при положительной.

В зависимости от типа переменных (количественные, ранговые, номинальные) используются различные методики нахождения зависимости: линейный коэффициент корреляции Пирсона, коэффициент ранговой корреляции Спирмена, коэффициент корреляции знаков Фехнера и др.

Например, коэффициент корреляции Спирмена вычисляется по формуле:

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

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:

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

где n – число одинаковых рангов в первом массиве, k – число одинаковых рангов во втором массиве [11].

Кластерный анализ

Кластерный анализ – статистические процедуры группировки объектов на основе мер сходства или расстояния между ними, выраженных численными значениями, и приемов (алгоритмов, правил) объединения объектов [12] (Впервые метод описан в работах [13–15]). Визуальным результатом анализа является дендрограмма, изображающая связи объектов в виде дерева.

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

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

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

[9] Корпус «Манускрипт». URL: manuscripts.ru.

[10] Корпус М. В. Ломоносова. URL: lomonosov.pro.

[11] Интернет-адрес модуля: http://manuscripts.ru/mns/!cred2.stat.

[12] Сейчас доступны Log-Likelihood, TF*ICTF, Weirdness.

Обоснование использования корреляционного и кластерного анализа

Установление степени близости двух или более текстов друг другу может быть осуществлено сопоставлением их различных как содержательных, так и формальных особенностей.

Одним из уникальных параметров любого текста является состав и количество лексических единиц, зависящих от жанра, регистра, тематики, стиля, времени и целей создания, индивидуальных сознательных и бессознательных предпочтений автора и некоторых других характеристик. Ярким примером характеризующих текст единиц являются жанрово и тематически значимые слова, отличающие его от текстов других жанров и тематики, – так называемые ключевые слова.

Несмотря на то, что перечень наиболее часто употребляемых в текстах слов (в первую очередь – служебных) примерно одинаков, их относительное количество может быть различно и поэтому может считаться индивидуальной характеристикой текста или группы текстов.

Это положение может быть экспериментально проверено сопоставлением состава и относительного количества наиболее частотных слов, в том числе и в средневековых славянских рукописях. Объективность результата может быть оценена, в частности, с помощью уже имеющихся сведений о жанре, типе, изводе анализируемых произведений.

Основой любого сопоставления является наличие в сравниваемых перечнях не только различных единиц, но и идентичных, а также не только различий в количественных значениях одной и той же единицы, но и совпадений. Соотношение совпадений и различий и позволяет выявить степень схожести (тесноты) или контраста сопоставляемых рядов лексем, а соответственно – текстов.

Таким образом, тексты могут быть представлены как массивы текстовых форм, отсортированных по частоте встречаемости или, например, по какому-либо статистическому критерию. Значениями элементов массива могут быть формы этих элементов, их абсолютное или относительное количество, статистическая величина, ранг – порядковый номер элемента в сортированном перечне. Эти значения и могут быть использованы для вычисления коэффициента корреляции (О корреляционном анализе как одном из перспективных методов для поиска статистических закономерностей в текстах, пишет в своей книге, вышедшей почти полвека назад, Б. Н. Головин [16]).

Визуализация результатов измерения расстояний между текстами может быть осуществлена различными способами. Одним из удобных приемов является построение дендрограммы на основе расстояний, значениями которых являются коэффициенты корреляции. Каждая из рукописей имеет свой набор расстояний по отношению к остальным. Соответственно, эти значения (включая значение 1,0 как отношение самого списка к себе) могут считаться переменными, на основе которых осуществляется группировка рукописей при кластерном анализе.

Машиночитаемые транскрипции корпуса «Манускрипт» как объект анализа

Транскрипции корпуса максимально точно передают графику рукописей: в текстовой форме сохраняются варианты букв, лигатуры, диакритика, титла. Поэтому при сопоставлении рядов текстовых форм нужно учитывать, что индивидуальные графико-орфографические особенности идентичных словоформ будут восприняты как дифференцирующие признаки, а оценка степени близости текстов будет во многом основана на графико-орфографических особенностях списков. Для нивелирования этого фактора в экспериментах должны быть предусмотрены разной степени унификация и нормализация форм, вплоть до использования для сопоставления перечней лемм.

Понятно, что состав находящихся в начале ранжированных по количественному или статистическому значению форм значительных по объему перечней разных текстов отличается: одно и то же слово имеет различные абсолютные и относительные частоты или статистические значения, соответственно в отсортированных списках имеет различные порядковые номера.

Закономерным следствием использования для сопоставления ранжированных по количественному или статистическому значению форм и для вычисления коэффициента корреляции их количественных или статистических значений, которые подчиняются закону Ципфа, так как принадлежат единицам текста, написанного на естественном языке, является отсутствие нормального распределения значений, а соответственно, возможность применения к рядам только ранговых корреляций, например ранговой корреляции Спирмена.

Процедура экспериментов

Все эксперименты выполнены по одному алгоритму:

1) в модуле статистики извлечение и ранжирование по количеству текстовых форм каждого из текстов в сопоставлении с теми же формами в других списках,

2) построение таблиц, содержащих ранги текстовых форм в различных текстах,

3) с помощью программы Statistica[13] вычисление коэффициента корреляции для каждой пары текстов (перечней),

4) построение сводных таблиц коэффициентов корреляции для всех пар текстов (перечней) и первоначальный анализ результатов,

5) с помощью программы Statistica на основе коэффициентов корреляции кластеризация всех рукописей, построение дендрограмм и повторный анализ результатов.

В качестве лингвистических единиц используются текстовые формы различной степени унификации.

[13] Statistica. Ver. 13. Copyright 1984-2017 TIBCO Software Inc. All rights.

Сопоставляются перечни наиболее частотных форм. Каждой единице перечня присваивается номер по порядку следования – ранг, при совпадении количественных значений единиц им присваивается один и тот же ранг.

Эксперимент 1: отношения списков XI века

Для всех текстов были построены перечни первых 100 наиболее частотных слов и форм[14] и для каждого элемента списка найдены ранги в других текстах, как это показано в табл. 1.

Таблица 1. **Наиболее частотные слова и текстовые формы в ЕО и их ранги в ОЕ, ЕА и МС[15]**

	ЕО			АЕ			МС		
Единицы	**№**	**R**	**F**	**№**	**R**	**F**	**№**	**R**	**F**
н	1	1	3985	1	1	2830	1	1	1858
къ	2	2	1638	2	2	1262	3	3	676
же	3	3	1193	3	3	835	7	7	359
не	4	4	946	4	4	644	22	21	144
отъ	5	5	826	435	101	11	91	62	42
ꙗко	6	6	695	7	7	437	4	4	428
рече	7	7	563	9	9	385	334	90	14
на	8	8	552	6	6	445	5	5	399
кмоу	9	9	507	10	10	381	874	98	6
кго	10	10	417	8	8	413	475	94	10
кстъ	11	11	403	12	12	270	108	68	36
да	12	12	395	11	11	295	48	42	68
къ	13	13	373	13	13	238	18	17	184
нїс	14	14	327	33055	112	0	34591	104	0
вамъ	15	15	314	16	16	195	2356	101	3
аще	16	16	287	17	17	192	649	96	8
азъ	17	17	283	21	21	158	537	95	9
глаа	18	18	279	36998	112	0	37995	104	0
имъ	19	19	263	14	14	207	880	98	6
съ	20	20	261	15	15	199	12	12	209
во	21	21	253	19	19	176	15	14	199
се	22	22	250	18	18	181	211	84	20
еваг	23	23	248	35068	112	0	36328	104	0
о	24	24	240	26	26	141	28	26	117
гъ	25	25	225	37304	112	0	123	71	33

Для каждой пары ранговых перечней найдены коэффициенты корреляции, как это показано в табл. 2.

[14] Значение параметра «Точность» – 0.5.2 (надстрочные буквы приравнены строчным, унифицированы, смещены в идентичные позиции титла, диакритика удалена).

[15] В таблице № – номер по порядку в сортированном по абсолютной частоте формы списке, R – ранг формы, F – абсолютное количество формы в тексте.

Таблица 2. **Коэффициенты корреляции для пар перечней (текстов) ЕО – [текст][16]**

ЕО	Valid N	R	p-value
ЕА	0,560909	6,707168	0,000000
МС	0,355825	3,769173	0,000280
МО	0,353391	3,739692	0,000310
МН	0,382419	4,097189	0,000086
МП	0,323586	3,385477	0,001023
ПЧ	0,449482	4,981193	0,000003
ПА	0,494058	5,625442	0,000000
И73	0,491385	5,585287	0,000000
И76	0,476350	5,363201	0,000001
13СГБ	0,474976	5,343213	0,000001
ПС	0,529518	6,179374	0,000000

Построена итоговая таблица, включающая коэффициенты корреляции для всех пар перечней (текстов) и цветом выделены близкие значения, как это показано в табл. 3.

Таблица демонстрирует статистическую близость (области крестов, или восклицательных знаков, или галочек) и удаленность текстов (области крестов vs галочек). Так, близкими являются два Евангелия, четыре минеи, Изборники, ПА и 13СГБ. Далекими Евангелия и минеи, минеи и ПЧ, ПА, И73. Степень близости и удаленности отмечается интенсивностью цветов.

Для визуализации близости рукописей построена дендрограмма (метод полной связи, Евклидово расстояние), как показано на рис. 1.

Таблица 3. **Коэффициенты корреляции для всех пар перечней (текстов)**

100	ЕО	ЕА	МС	МО	МН	МП	ПЧ	ПА	И73	И76	13СГБ	ПС
ЕО	1,000	0,561	0,356	0,353	0,382	0,324	0,449	0,494	0,491	0,476	0,475	0,530
ЕА	0,561	1,000	0,406	0,437	0,362	0,384	0,451	0,475	0,444	0,431	0,443	0,472
МС	0,356	0,406	1,000	0,572	0,515	0,314	0,175	0,204	0,184	0,145	0,165	0,283
МО	0,353	0,437	0,572	1,000	0,455	0,528	0,274	0,245	0,297	0,295	0,222	0,340
МН	0,382	0,362	0,515	0,455	1,000	0,466	0,280	0,340	0,353	0,349	0,316	0,373
МП	0,324	0,384	0,314	0,528	0,466	1,000	0,312	0,323	0,354	0,342	0,306	0,336
ПЧ	0,449	0,451	0,175	0,274	0,280	0,312	1,000	0,613	0,599	0,616	0,569	0,556
ПА	0,494	0,475	0,204	0,245	0,340	0,323	0,613	1,000	0,699	0,671	0,746	0,546
И73	0,491	0,444	0,184	0,297	0,353	0,354	0,599	0,699	1,000	0,739	0,688	0,597
И76	0,476	0,431	0,145	0,295	0,349	0,342	0,616	0,671	0,739	1,000	0,725	0,655
13СГБ	0,475	0,443	0,165	0,222	0,316	0,306	0,569	0,746	0,688	0,725	1,000	0,596
ПС	0,530	0,472	0,283	0,340	0,373	0,336	0,556	0,546	0,597	0,655	0,596	1,000

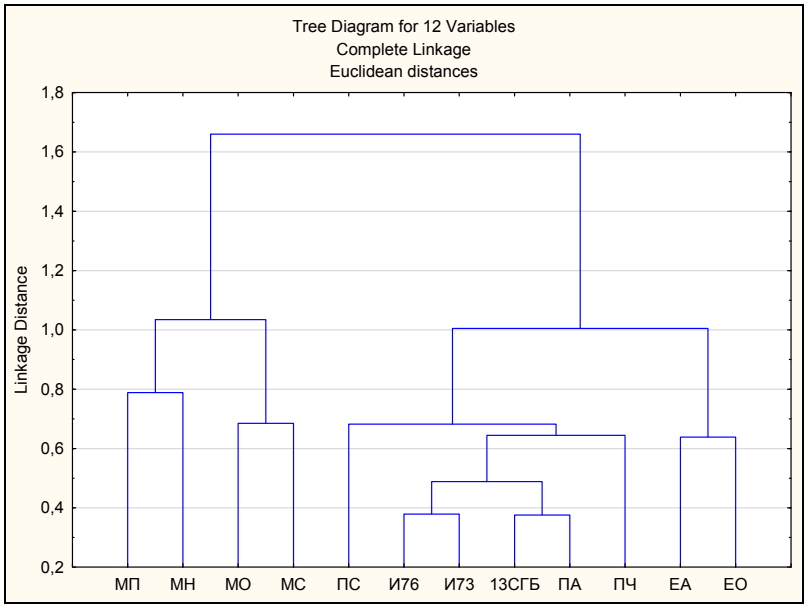

Рис. 1. Дендрогамма 12 рукописей XI века на основе текстовых форм (100 наиболее частотных, условная форма 0.5.2)

Рисунок позволяет увидеть группировку (кластеры) рукописей и отношения между кластерами разных уровней. На уровне двух кластеров минеи противопоставляются всем другим рукописям, на уровне трех – минеи vs Евангелия vs сборники и переводные сочинения различных авторов. В группе миней выделены две подгруппы – МП и МН, с одной стороны, и МО и МС, с другой, в группе разножанровых рукописей Патерик противопоставлен всем другим текстам. Видны и более частные группировки.

Понятно, что на группировку рукописей влияет несколько факторов: состав перечней (зависит от

[16] Valid N – количество совпавших форм, R – коэффициент корреляции, p-value –достоверность результата (при значении меньше 0,05 результат достоверен).

критерия отбора и порядка следования), ранги форм в перечнях (зависят от относительного количества форм в текстах), количество идентичных форм (зависит от наличия формы в других текстах). В свою очередь эти три параметра зависят от степени унификации текстовых форм, извлекаемых при выборке, и от количества единиц выборки. При использовании оригинальных форм транскрипции, с сохранением в них диакритики, титл, их позиции, вариантов букв и др., ожидается большая вариативность перечней и, соответственно, большее влияние на результат измерений фактора написания. При использовании унифицированных форм – меньшее влияние этого фактора. При увеличении количества форм в выборке, например со ста до трехсот, ожидается большее их разнообразие, которое может оказывать существенное влияние на степень тесноты перечней (текстов).

Для установления влияния на группировку списков двух параметров выборок: степени унификации и количества единиц в выборке – были осуществлены еще 11 аналогичных извлечений и измерений на основе перечней в разной степени унифицированных текстовых форм (от форм без унификации до форм, в которых удалены титла, диакритика, лигатуры раскрыты) объемом по 50, 100, 200 и 300 единиц. Результаты корреляционного сопоставления текстов показаны на рис. 2.

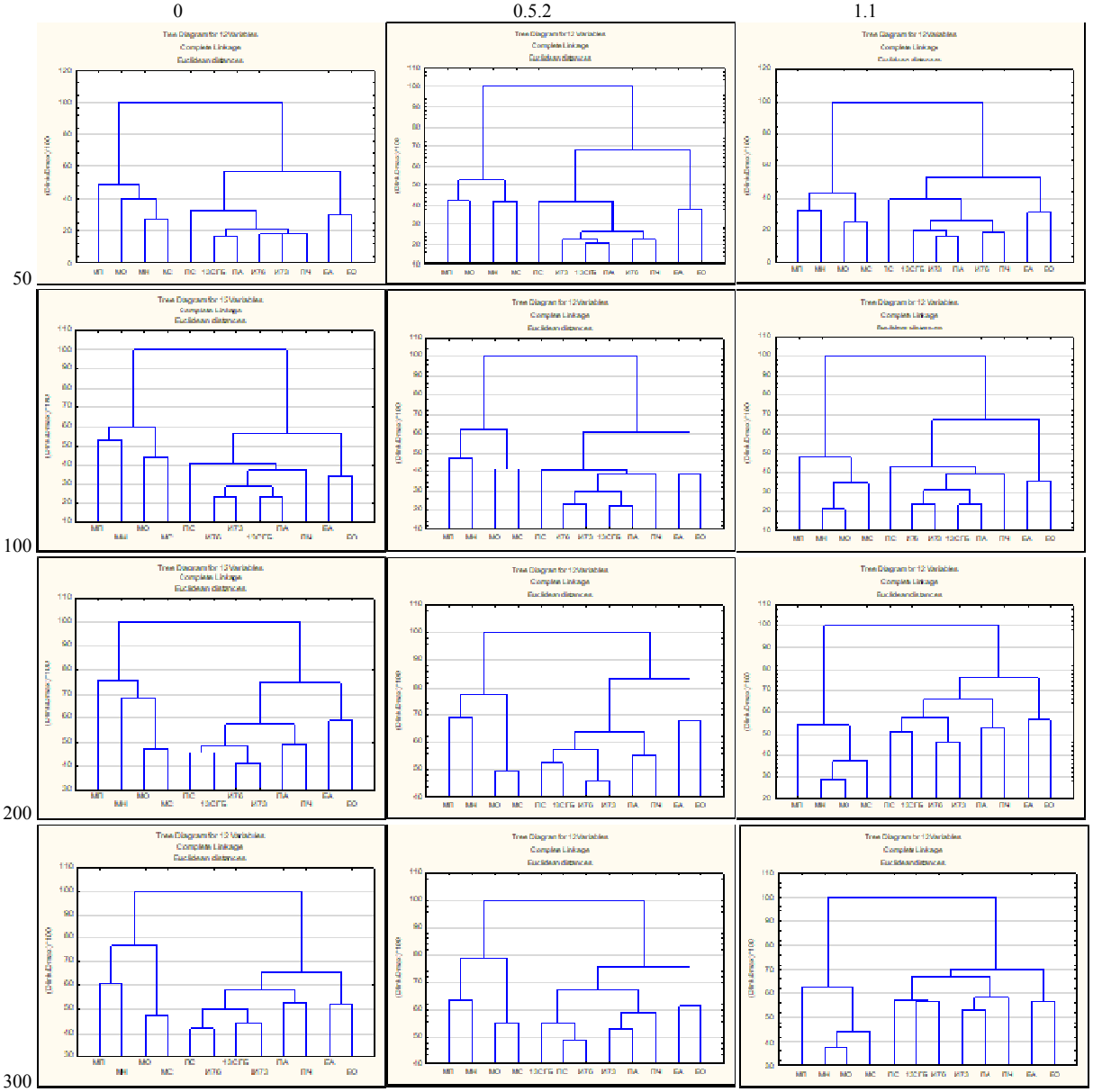

Рис. 2. Дендрограммы рукописей XI века на основе текстовых форм (50, 100, 200 и 300 наиболее частотных, условные формы 0, 0.5.2, 1.1)

Дендрограммы демонстрируют наличие трех кластеров (минеи и Евангелия противопоставлены всем другим текстам) и их различную устойчивость при изменении параметров выборки: наличие трех вариантов кластера миней (кластер А), разнообразие конфигурации кластера сборников различного со-

держания (кластер В имеет пять вариантов соотношения списков) (см. табл. 4а и 4б), стабильность кластера двух Евангелий (кластер С).

Табл. 4а, 4б: **Варианты конфигураций кластеров А и В**[17]

	Кластер А			**Кластер В**		
		Унификация			*Унификация*	
Выборка	*0*	*0.5.2.*	*1.1.*	*0*	*0.5.2.*	*1.1.*
50	**1**	**3**	**3**	**2**	**1А**	**1Б**
100	**3**	**3**	**2**	**3**	**3**	**3**
200	**1**	**3**	**2**	**4А**	**4А**	**4А**
300	**3**	**3**	**2**	**4Б**	**5А**	**5Б**

В группе миней (кластер А) в 7 экспериментах из 12 попарно близкими являются МС-МО и МН-ПМ (конфигурация 3, табл. 4а), в пяти – МП противопоставлена трем другим минеям (конфигурации 1 и 2). Кластер оказался стабильным при средней и максимальной степени унификации.

Варианты конфигурации кластера В (табл. 4б):

– наиболее часто зафиксирован вариант 4А, в котором ПА и ПЧ противопоставлены попарно группирующимся И73-И76 и 13СГБ-ПС: (((И73 И76) (13СГБ ПС)) (ПЧ ПА)),

– один раз образуется близкий к предыдущему вариант 4Б: (((И73 13СГБ) (И76 ПС)) (ПЧ ПА));

– в трех случаях (конфигурация 3) последовательно ПС и ПЧ противопоставлены группам И73-И76 и ПА-13СГБ: ((((И73 И76) (ПА 13СГБ)) ПЧ) ПА);

– в двух случаях (конфигурация 1) кластер В разбивается на три подкластера:

1А – (((И76-ПЧ) (И73 (13СГБ ПА))) ПС) и

1Б – (((И76-ПЧ) (13СГБ (И73 ПА))) ПС);

– также в двух случаях (конфигурация 5) образуются два подкластера:

5А – (((И76 ПС) 13СГБ) (ПА И73) ПЧ))) и

5Б – (((И76 13СГБ) ПС) (ПА И73) ПЧ)));

– один раз (конфигурация 2) подкластеры выглядят следующим образом – ((((И76 И73) ПЧ) (13СГБ ПА)) ПС).

На изменение конфигурации кластера В влияет объем выборки. В то же время конфигурации выборок из 100 и 200 единиц (конфигурации 3 и 4) достаточно стабильны при изменении степени унификации.

Учет конфигураций всех кластеров Евангелий, миней и сборников позволяет увидеть стабильность дендрограммы при выборке 50 единиц и средней и максимальной степени унификации. В то же время 6 текстов кластера В показывают устойчивость конфигурации к изменению степени унификации при выборке 100 и 200 единиц.

Эксперимент 2. Произведения Кирилла Туровского в отношении к текстам списков XI века

Результаты первого эксперимента позволяют поставить вопрос о степени близости и других текстов к выявленным группам рукописей XI века и подтвердить или отвергнуть текстолого-лингвистическое предположение, что, например, авторский текст средневекового русского проповедника будет более близок группе сборников, чем миней или Евангелий.

Для проверки подобной гипотезы была исследована степень близости кодексов XI в. рукописи, содержащей произведения автора вт. пол. XII в. Кирилла Туровского в списке, наименее удаленном от оригинала. Гомилетические сочинения Кирилла Туровского отстоят от времени создания списков старославянских переводов на столетие или менее, чем столетие. Эти списки демонстрируют исходную точку формирования древнерусского извода церковнославянского языка. Они вместе с тем представляют значимый срез канонической литературы, которая является ядром литературного процесса в Древней Руси [17]. Греческие оригиналы этих текстов – прежде всего учительные – для средневековых славянских писателей были источниками стилистических и концептуальных клише [18]. В проповедях Кирилла Туровского влияние византийских образцов сказалось наиболее ярко, и в период его деятельности развитие литературного языка достигает наивысшей точки в домонгольской Руси.

Извлечение и сопоставление рангов первых 100 наиболее частотных словоформ СбТол с их рангами во всех текстах XI в. (см. начало этого списка в табл. 5), построение таблицы корреляции СбТолКТ с каждой из сопоставляемых рукописей (см. пример в табл. 6), добавление в табл. 6 значений для каждой пары текстов и построение дендрограммы для выборки в 100 словоформ (унификация 0.5.2) (рис. 3), а затем построение дендрограмм на основе разного объема выборок (50, 100, 200, 300 единиц) с разной степенью точности словоформ (степени унификации 0, 0.5.2, 1.1) (см. рис. 4) позволили наглядно представить, каким образом количество анализируемых словоформ и степень их точности влияют на близость анализируемых текстов Кирилла Туровского текстам списков XI века.

[17] Условным номером обозначен вариант соотношения списков в каждом кластеров.

Таблица 5. **Ранги словоформ 12 рукописей XI века в сопоставлении с рангами наиболее частотных словоформ текстов Кирилла Туровского**[18]

Единицы	СбТолКТ	ЕО	ЕА	МС	МО	МН	МП	ПЧ	ПА	И73	И76	13СГБ	ПС
н	1	1	1	1	1	1	1	1	1	1	1	1	1
бъ	2	2	2	3	3	3	2	4	4	3	4	4	3
на	3	8	6	5	6	8	5	7	9	8	9	9	9
не	4	4	4	21	18	19	24	6	3	4	3	3	4
же	5	3	3	7	9	9	10	2	6	5	6	6	2
ѿ	6	44	5	10	13	11	18	101	95	98	81	81	11
бо	7	21	19	14	24	18	16	5	5	7	5	5	28
съ	8	20	15	12	10	14	9	25	15	25	11	11	19
нъ	9	29	30	59	57	63	55	12	13	13	20	20	25
него	10	10	8	94	63	62	32	17	39	29	14	14	15
нако	11	6	7	4	5	4	3	8	7	22	7	7	6
да	12	12	11	42	47	31	30	18	10	9	12	12	10
къ	13	13	13	17	17	15	14	21	21	15	13	13	8
ѡ	14	141	112	104	90	96	75	101	131	146	82	82	127
по	15	26	25	29	23	40	42	14	22	14	24	24	16
кость	16	11	12	68	29	57	49	11	17	10	8	8	26
бса	17	141	82	67	37	33	34	101	55	66	75	75	72
ксн	18	42	41	16	4	5	6	19	106	112	45	45	57
а	19	39	33	94	85	92	71	48	16	27	21	21	90
нн	19	50	40	66	62	76	65	23	20	19	19	19	32
ан	20	38	27	98	83	95	69	82	25	26	22	22	38
нъна	21	141	112	104	90	93	73	101	119	139	81	124	120
та	22	67	55	11	7	6	4	51	60	69	28	88	84
аще	23	16	17	96	81	92	72	101	23	71	36	16	33
бꙑ	23	141	112	104	90	96	11	101	130	146	82	125	127
нъна	21	141	112	104	90	93	73	101	119	139	81	124	120

[18] Показаны первые 25 словоформ; значение параметра «Точность» – 0.5.2.

Таблица 6. **Коэффициенты корреляции для пар перечней (текстов) СбТолКТ – [текст]**[19]

СбТолКТ	Valid N	R	p-value
ЕО	0,564665	6,773004	0,000000
ЕА	0,542369	6,390809	0,000000
МС	0,430662	4,723843	0,000008
МО	0,451667	5,011595	0,000002
МН	0,417188	4,544299	0,000016
МП	0,480483	5,423627	0,000000
ПЧ	0,452809	5,027520	0,000002
ПА	0,578719	7,024932	0,000000
И73	0,518395	6,001165	0,000000
И76	0,538378	6,324487	0,000000
13СГБ	0,517522	5,987357	0,000000
ПС	0,558050	6,657460	0,000000

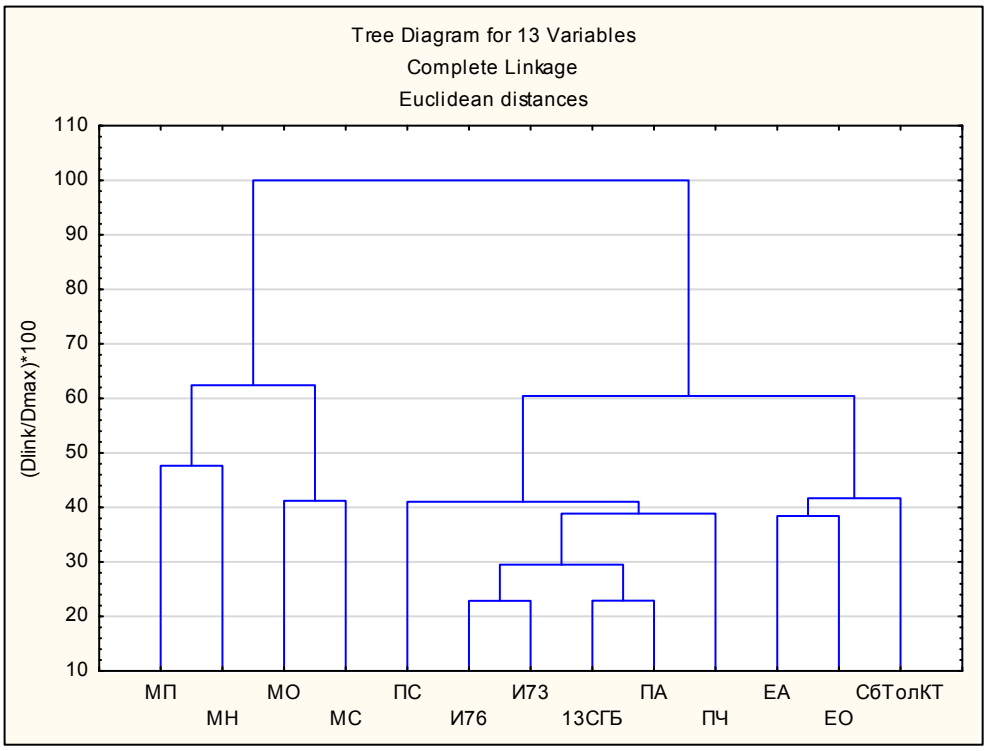

Рис. 3. Дендрограмма рукописей XI века и СбТолКТ XIII века на основе 100 наиболее частотных текстовых форм (условная форма 0.5.2)

Дендрограммы показывают, что в зависимости от объема выборки и степени унификации словоформ СбТолКТ сближается с разными группами текстов: в 4 случаях СтТолКТ включается в группу Евангелий, в 4 – в кластер миней, в 4 – в группу сборников.

В зависимости от объема выборки и степени точности словоформ эти сближения распределяются так, как показано в табл. 7. Кластеризация перечней словоформ без их графико-орфографической унификации (условная форма 0) последовательно сближает СбТолКТ с группой миней. При средней (условная форма 0.5.2) и высокой (условная форма 1.1) степени унификации при больших выборках (200 и 300 единиц) СбТолКТ входит в кластер сборников, при меньших выборках (50 и 100 единиц) – в кластер Евангелий.

[19] Для выборки 100 словоформ.

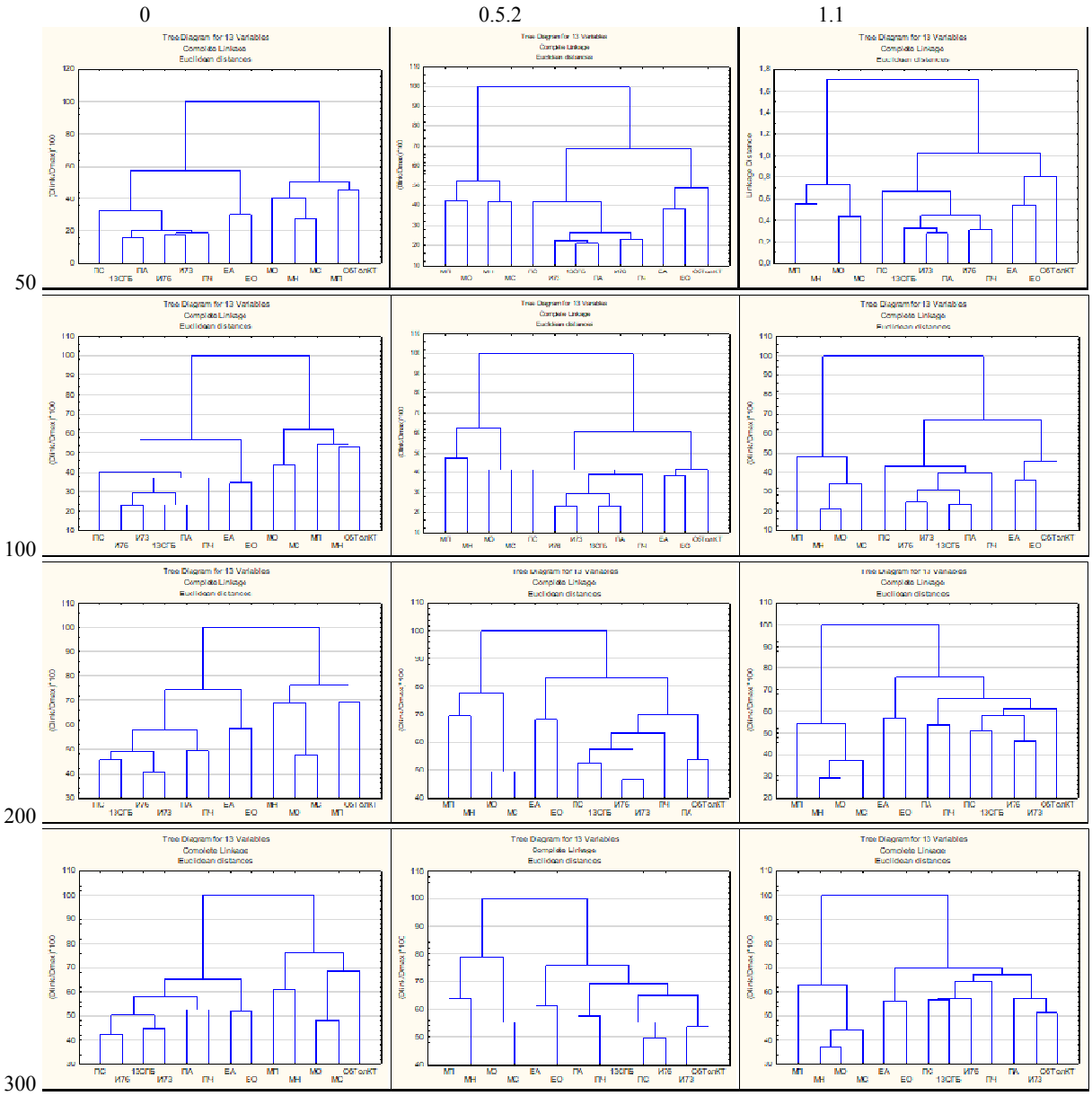

Рис. 4. Дендрограммы рукописей XI века и СбТолКТ на основе текстовых форм (50, 100, 200 и 300 наиболее частотных, условные формы 0, 0.5.2, 1.1)

Таблица 7. **Вхождение СбТолКТ в кластеры**[20]

	Унификация		
Выборка	*0*	*0.5.2.*	*1.1.*
50	**1А**	**2**	**2**
100	**1Б**	**2**	**2**
200	**1В**	**3А**	**3В**
300	**1Г**	**3Б**	**3Г**

В группе миней четыре конфигурации:
1А – (СбТолКТ МП) (МО (МС МН)),
1Б – ((СбТолКТ МН) МП) (МС МО),
1В – (СбТолКТ МП) (МН (МС МО)),
1Г – (СбТол (МС МО)) (МН МП).

В группе 2 (Евангелия – СбТолКТ) соотношение списков не меняется: (СбТолКТ (ЕО ЕА)).

В группе сборников их также четыре:
3А – (СбТолКТ ПА) (((И73 И76) (13СГБ ПС)) ПЧ),
3Б – ((СбТолКТ И73) ((И76 ПС) 13СГБ)) (ПЧ ПА),
3В – (СбТолКТ ((И73 И76) (13СГБ ПС))) (ПЧ ПА),
3Г – ((СбТолКТ И73) ПА) (((13СГБ ПС) И76) ПЧ).

[20] Условные номера: 1 – СбТолКТ находится в кластере миней, 2 – в кластере Евангелий, 3 – в кластере сборников.

Сближаясь с Евангелиями, СбТолКТ противопоставлен обоим спискам. В кластере миней СбТолКТ образует подкластеры с разными списками: 2 раза с МП (конфигурации 1А, 1В), один раз с МН (1Б), один раз с парой (МС МО) (1Г). В кластере сборников СбТолКТ оказывается наиболее близок ПА (3А), И73 (2 раза) (3Б, 3Г), а также попарно сгруппированным Изборникам и 13СГБ и ПС (3В).

Обсуждение экспериментов

Наиболее высокий уровень унификации графики текстовых форм дает довольно точное и полное представление о тесноте лексико-грамматических и лексико-семантических отношений между текстами.

Горизонтальный уровень перехода степеней унификации вместе с тем указывает на возможность статистической оценки близости графико-орфографического оформления рукописей.

Вертикальный ряд статистических данных, в котором представлены разные конфигурации близости текстов в зависимости от охвата наиболее частотных единиц – 50, 100, 200 и 300, – как показывает анализ, при наибольшем уровне унификации характеризует три уровня близости текстов. Уровень 50 и 100 языковых единиц по составу связан прежде всего с частотностью собственно грамматических, служебных лексем – предлогов, обеспечивающих внутрисинтагменные связи, частиц, союзов и союзных слов, необходимых в синтаксисе клауз, а также словоформ дейктических лексем, обслуживающих тестовую связность – местоимений разных семантических разрядов, местоименных энклитик, связочных форм глагола существования **быти**. Употребление всех названных единиц в большой степени обусловливается речевым автоматизмом. На уровне данных единиц проповеди Кирилла Туровского обнаруживают наибольшую близость безыскусному языку Евангелия, а статистический анализ обнаруживает скрытый потенциал их речевого воздействия. По рангу частотности языку Евангелия здесь оказываются близки следующие единицы СбТолКТ – в порядке уменьшения частотности в каждом типе единиц[21]: а) предлоги **въ**, **на**, **ѿ**, **съ**, **къ**, **по** и послелог **ради**; б) союзные средства **же**, **бо**, **нъ**, **ꙗко**, **да**, **а**, **ли**, **аще**, **иже**, **ꙗкоже**; в) отрицательная и усилительная частицы **не**, **ни**; г) словоформы личных и возвратного местоимений, а также соотносительных с ними притяжательных местоимений **та**, **ми**, **ма**, **ти**, **ты**, **са**, **си**, **вамъ**, **азъ**, **мнѣ**, **вы**, **тебе**, **мои**, **своꙗ**; д) словоформы анафорического, указательных и вопросительного местоимений **ѥго**, **се**, **ѥмоу**, **сего**, **тъ**, **кто**; е) местоименное наречие **тако**; ж) глагол существования **быти** в форма презенса ед. ч. **ѥсть**, **ѥси**, **ѥсмь** и аориста 3 л. ед. **бѣ**. Обращает на себя внимание наличие в этом перечне послелога **ради**, свойственного кирилло-мефодиевской традиции и противопоставленного преславскому и древнерусскому **дѣлꙗ**, отсутствие преславского сравнительного союза **акы** при наличии кирилло-мефодиевского **ꙗко** (см. [19]). Есть лишь четыре показательных исключения в статусе лексем – это ключевые для христианских текстов субстантивные словоформы **бога**, **сынъ**, а также аорист **рече** и причастие **глагола**, вводящие цитирование и прямую речь, определяющие, таким образом, драматургическую выразительность и насыщенность проповедей Кирилла Туровского вслед за евангельскими текстами. Эти формы значимы как средства введения «тематических ключей» при цитировании или аллюзий на канонические тексты. Ориентированность на текст Евангелия подчеркивается уже в названиях проповедей: **того жⷤ · кюрила · мниха · слⷡо · [..] сънатии тѣла хⷭва съ крⷭта · и [ѡ] мюроносицⷯ а · ѿ сказаниꙗ ѥвангⷧскааго · и похвала иѡсифоу · въ нⷣ е · г҃ю по пасцѣ⁘** СбТолКТ, 5.2–6.1; **кюрила · мнⷯи · слⷡо · [о] слѣпьци · и [ѡ] зависти жидовъ · ѿ сказаниꙗ ѥоуⷢльскаго · в нⷣ е · г҃ю по пасцѣ · гⷭ и блгⷭви ѡтче⁘** 25.1 и др.

Резкое изменение конфигурации наиболее близких текстов происходит на уровне 200 частотных слов или словоформ. СбТолКТ в этом случае оказывается близким двум подкластерам – И73 и И76, а также 13СГБ и ПС. Слова и словоформы с рангом от 1 до 200 отражают восполнение и расширение того же круга единиц, которые представлены на первом уровне – с рангом с 1 до 100. Укажем близкие по рангу единицы в данных текстах этого типа:

[21] Указываются единицы с рангом до 100 в одном из Евангелий. Опускается союз или местоимение **и**, которое имеет первый ранг во всех текстах.

а) предлог **до**; б) союзные средства **кгда**, **нлн**, **нхъже**, **оубо**, **понеже**; в) формы личных и притяжательных местоимений **ны**, **мене**, **твоѭ**, **васъ**, **мы**, **своемь**, **свон**, **свонхъ**; г) формы анафорического, указательного и вопросительного местоимений **ѹто**, **нхъ**, **ѭ**, **нмъ**, **ннмь**, **томоу**, **тѣхъ**; д) наречия **оуже**, **прѣже** (употреблялось и в функции предлога); е) экзистенциальный глагол **бытн** в формах наст. времени 3 л. мн. ч. **соуть**, наст. времени сов. вида 3 л. ед. ч. **боудеть**, аориста 1 л. и 3 л. ед. ч., входящего в состав форм условного наклонения **быхъ**, **бы**, при том что известно и самостоятельное их предикативное употребление. Наряду с этим значительно расширяется ряд лексем христианского топикального свода, входящих в так называемые тематические ключи: **інсоусъ**, **мнръ**, **слово**, **божнн**, **зємлю**, **прнде**, **сѧтаго**, **цьркъвн**, **ѹелов҃ѣкъ**, **законъ**, **роуцѣ**, **цьркъвь**, **доуха**, **прнсно**. Здесь встречаются лексемы и с нейтральной тематической принадлежностью, которые употребляются вне книжных текстов, – такие, как аорист основного глагола движения **прнде** или антропометрическое обозначение **роуцѣ**, однако в данном случае они также входят в контексты вероисповедной направленности.

Довольно существенно меняется конфигурация распределения текстов на следующем уровне – с рангом частотности до 300. Близость языку «славянской энциклопедии» – И73 – выступает еще более отчетливо, чем на втором уровне. СбТолКТ образует пару с И73, которая, в свою очередь, входит в подкластер с ПА. Этот качественный сдвиг обеспечивают прежде всего следующие лексемы и формы с рангом частотности до 300: а) предлог **прѣдъ**; б) союзные средства **ндеже**, **неже**; в) формы притяжательных местоимений **твон**, **твоего**, **мое**, **своемъ**; г) формы анафорического, указательного и определительного местоимений **снн**, **вьсь**, **немь**, **ннхъ**, **снхъ**, **всего**; д) местоименные наречия **снце**, **тъгда**. За счет использования тематических ключей расширяется круг собственно знаменательных лексических единиц: **нзынѣ**, **свѣтъ**, **слава**, **дабыдъ**, **дѣла**, **жнвотъ**, **молю**, **сыноу**, **сы**, **ѥдннъ**, **вндѣ**, **врѣма**, **дьнь**, **добрѣ**. Близость текстов вновь конституируется расширением круга двух лексических рядов, при том что первый перечень становится менее разнообразным и объемным, а во втором перечне обнаруживается ветхозаветный персонаж, но преобладают единицы с нейтральной тематической принадлежностью – в частности, аорист основного глагола зрительного восприятия **вндѣ**, а также такие лексемы, как **нзынѣ**, **ѥдннъ**, **дѣла**, **дьнь**, **добрѣ**, которые тем не менее оказываются соотнесены с вероисповедной направленностью текстов. Обращает на себя появление в этом ряду сугубо книжной причастной формы **сы**, связанной с исповеданием единственно «сущего».

Таким образом, анализ языковых единиц, частотность которых определяет близость текстов, в случае с проповедями Кирилла Туровского, наряду с ожидаемыми выводами, дает и парадоксальный результат. Близость текстов имеет иерархический, трехуровневый характер. Наибольшая близость проповедей языку Евангелия проявляется на уровне 50 и 100 частотных единиц и определяется не тематически, а структурно – единицами внутрисинтагменной и клаузной связи, а также дейктиками, связочными глагольными формами и глагольными форами, вводящими цитирование. Таким образом, теснота текстового сходства выступает в этом случае подспудно, имплицитно и соотносится с речевым автоматизмом употребления и восприятия. На уровне 200 единиц проявляется близость СбТолКТ и двух сборниковых подкластеров – И73 и И76, 13СГБ и ПС. Это сближение носит комбинированный, структурно-тематический характер: наряду с пополнением круга союзных средств, дейктиков и форм глагола **бытн**, существенно расширяется круг тематически значимых единиц, связанных с обращением к тематическим ключам вероисповедных текстов. На уровне 300 единиц яснее обнаруживается близость СбТолКТ и сборников экзегетической направленности – И73 и ПА. При незначительной роли синтаксических маркеров и сохранении значимости дейктических единиц она проявляется главным образом тематически – репертуаром полнозначных лексических единиц, среди которых есть такие тесно связанные с вероисповедной направленностью, как **свѣтъ**, **слава**,

дабъıдъ, **жнвотъ**, **молю**, **сънѹ**, **съı** ‘сущий’, с одной стороны, и тематически нейтральные – такие, как **нъınѣ**, **дѣла**, **кдннъ**, **вндѣ**, **врѣма**, **дьнь**, **добрѣ**, с другой стороны. Безусловно, в контекстах вероисповедной направленности тематическая нейтральность лексем стирается, но они тем не менее выстраивают своего рода мост между текстами обычного и экзегетического типов.

Выводы

Создание больших коллекций размеченных машиночитаемых копий славянских средневековых письменных памятников позволяет использовать их для получения количественных и статистических сведений о лингвистических единицах. Наличие у информационного ресурса инструментов для извлечения данных на основе мета-, аналитической и лингвистической разметок позволяет получить материал для сопоставительного анализа количественных характеристик нескольких выборок.

Сопоставление рангов – мест идентичных словоформ в ранжированных перечнях – наиболее частотных слов 12 древнейших русских рукописей с помощью ранговой статистики Спирмена позволяет установить степень тесноты каждой пары текстов, а использование метода кластеризации – группировку всех кодексов. Полученные результаты – группы Евангелий, миней, сборников – отражают жанрово-типовые характеристики текстов и демонстрируют различное соотношение списков внутри групп в зависимости от объемов выборок и степени унификации текстовых форм. Объем сопоставляемых перечней и степень соответствия единиц выборок их исходной графической форме существенно влияют и на место текстов Кирилла Туровского в дендрограммах: или в группе миней, или сборников, или Евангелий.

Лингвистическая интерпретация дендрограмм, полученных на основе максимально унифицированных текстовых форм, позволила выявить зависимость сближения текстов Кирилла Туровского с разными группами текстов XI века от лексико-грамматического и лексико-семантического состава сопоставляемых перечней разного объема. Лингвостатистический анализ выявил три уровня близости текстов – структурный, структурно-тематический и преимущественно тематический. СбТолКТ обнаруживает три уровня текстовых сближений. Если на первом уровне близость статистически определяется не только и не столько лексемами христианского топикального круга, сколько разнообразием и полнотой использования единиц служебного характера, на втором уровне она осложняется единицами тематических ключей, а на третьем уровне лексически значимые схождения играют главенствующую роль. Лингвостатистический анализ доказывает, что иерархия текстовых схождений обусловливает коммуницирующую и аргументативную действенность проповедей Кирилла Туровского. Искусство проповеди основывается на полновесном использовании механизма интертекстуальности, и проведенный анализ подтверждает, что Кирилл Туровский опирался на него.

Источники

ЕО – Евангелие апракос краткий (Остромирово Евангелие), 1056–1057 г., РНБ, F.п.1.5., 294 л. СК № 3.

ЕА – Евангелие апракос краткий (Архангельское евангелие), 1092 г., РГБ, М.1666, 178 л. СК № 6.

И73 – Изборник 1073 г., ГИМ, Синод. 1043, 266 л. СК № 4.

И76 – Изборник 1076 г., РНБ, Эрм. 20, 277 л. СК № 5.

МС – Минея служебная на сентябрь, 1095–1096 г., РГАДА, ф. 381 (Син.тип.) № 84, 176 л. СК № 7.

МО – Минея служебная на октябрь 1096 г., РГАДА, ф. 381 (Син.тип.) № 89, 127 л. СК № 8.

МН – Минея служебная на ноябрь 1097 г., РГАДА, ф.381 (Син.тип.) № 91, 174 л. СК № 9.

МП – Служебная минея на май (Путятина Минея), XI в., РНБ, Соф. 202, 135 л. СК № 21.

ПА – Пандекты черноризца Антиоха, XI в., ГИМ, Воскр. 30, 309 л. СК № 24.

ПЧ – Чудовская псалтырь, XI в., ГИМ, Чуд. 7, 176 л. СК № 31.

13СГБ – 13 слов Григория Богослова, XI в., РНБ, Q.п.1.16, 377 л. СК № 33.

ПС – Синайский патерик, XI в., ГИМ, Синод. 551, 184 л. СК № 26.

СбТолКТ – Сборник Слов и поучений («Толстовский сборник»), втор. пол. XIII в. (РНБ, F.п.I. 39), 184 л. [Электронный ресурс] / О. Ф. Жолобов и др.; проект «Манускрипт». URL: http://manuscripts.ru/mns/main?p_text=96362255 (дата обращения: 08.08.2019).

Библиографические ссылки

1. *Головин Б. Н.* Язык и статистика. М. : Просвещение, 1971. 190 с.

2. Там же. С. 157–159.

3. Компьютеризованный статистический анализ для историков / под ред. Л. И. Бородкина и И. М. Гарсковой. М., 1999. 187 с.

4. *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.

5. *Victor Baranov*. A Text Corpus of Medieval Manuscripts as a Goal and a Tool for Linguistic Research // Editing Mediaeval Texts from a Different Angle: Slavonic and Multilingual Traditions (together with Francis J. Thomson’s Bibliography and Checklist of Slavonic Translations). To Honour Francis J. Thomson on the Occasion of His 80th Birthday : Together with Proc. of the ATTEMT Workshop held at King’s College, London, 19–20 December 2013 and the ATTEST Workshop held at the University of Regensburg, 11–12 December 2015 / edited by Lara Sels, Jürgen Fuchsbauer, Vittorio Tomelleri and Ilse de Vos. Peeters Leuven - Paris - Bristol, Ct, 2018. Pp. 283-308.

6. *Баранов В. А.* Поиск и демонстрация данных в историческом корпусе «Манускрипт» // Корпусная лингвистика –2019 : труды международной конференции (24–28 июня 2019 г., Санкт-Петербург). СПб. : Изд-во С.-Петерб. ун-та, 2019. С. 271–279.

7. *Баранов В. А., Дубовцев С. В.* Модуль статистики информационно-аналитической системы «Манускрипт»: функции и демонстрация данных // Информационные тех-

нологии и письменное наследие: материалы IV Междунар. науч. конф. (Петрозаводск, 3–8 сентября 2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. Петрозаводск ; Ижевск, 2012. С. 23–26.

8. Сводный каталог славяно-русских рукописных книг, хранящихся в СССР (XI–XIII вв.). М. : Наука, 1984. 406 с.

9. *Баранов В. А., Жолобов О. Ф.* Лингвостатистическое исследование частотных слов в Словах Кирилла Туровского (по рукописи РНБ, F.п.I.39) // Slověne = Слов□не. International Journal of Slavic Studies. В печати.

10. *Жолобов О. Ф.* О контрастирующих орфографических системах в рукописи XIII в. (к интернет-изданию Толстовского сборника) // Древняя Русь. Вопросы медиевистики. 2018. 3 (73). С. 77–89.

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

12. *Paul A. and Jr. Gore.* Cluster analysis. In: Handbook of Applied Multivariate Statistics and Mathematical Modeling. (Eds.) Howard E.A. Tinsley and Steven D. Brown. Academic Press, 2000. Pp. 297-321.

13. *Tryon, R.* Cluster analysis. New York: McGraw Hill, 1939.

14. *Cattell, R. B.* A note on correlation clusters and cluster search methods. Psychometrica, 9, 1944. Pp. 169-184.

15. *Sokal, R. and P. Sneath.* Principles ofnumeric taxonomy. San Francisco: W. H. Freeman, 1963.

16. *Головин Б. Н.* Указ. соч. С. 159–166.

17. Успенский 1988 – *Успенский Б. А.* История русского литературного языка (XI–XVII вв.). Budapest: Tankönyvkiadó, 1988. 451 с. С. 18, 68.

18. *Picchio, R.* Models and patterns in the literary tradition of Medieval Orthodox Slavdom // American contributions to the Seventh International Congress of Slavists, II. The Hague, 1973. P. 445.

19. *Пичхадзе А. А.* Переводческая деятельность в домонгольской Руси: лингвистический аспект. М.: НП «Рукописные памятники Древней Руси», 2011. 408 с. С. 54.

References

1. Golovin B. N. *Yazyk i statistika* [Language and statistics]. Moscow : Prosveshchenie, 1971, 190 p. (in Russ.).

2. Ibid. Pp. 157-159.

3. *Komp'yuterizovannyi statisticheskii analiz dlya istorikov* [Computerized Statistical Analysis for Historians] (eds. Borodkin L.I., Garskova I.M.). Moscow, 1999, 187 p. (in Russ.).

4. Baranov V. A. [The historical corpus as the goal and instrument of corpus paleoslavistics]. *Scripta & e-Scripta* : The Journal of Interdisciplinary Mediaeval Studies. Sofia : “Boyan Penev” Publishing Center ; Institute of Literature, BAS, 2015, vol. 14-15, pp 39-62. (in Russ.).

5. Victor Baranov. *A Text Corpus of Medieval Manuscripts as a Goal and a Tool for Linguistic Research*. Editing Mediaeval Texts from a Different Angle: Slavonic and Multilingual Traditions (together with Francis J. Thomson’s Bibliography and Checklist of Slavonic Translations). To Honour Francis J. Thomson on the Occasion of His 80th Birthday : Together with Proc. of the ATTEMT Workshop held at King’s College, London, 19-20 December 2013 and the ATTEST Workshop held at the University of Regensburg, 11–12 December 2015 (eds. Lara Sels, Jürgen Fuchsbauer, Vittorio Tomelleri and Ilse de Vos). Peeters Leuven - Paris - Bristol, Ct, 2018, pp. 283-308. (in Eng.).

6. Baranov V. A. *Poisk i demonstratsiya dannykh v istoricheskom korpuse «Manuskript»* [Search and data demonstration in the historical corpus "Manuscript"]. *Trudy mezhdunarodnoi konferentsii «Korpusnaya lingvistika–2019»* (24-28 iyunya 2019 g., Sankt-Peterburg) [Proc. of the international conference “Corpus Linguistics –2019” (June 24–28, 2019, St. Petersburg)]. St. Petersburg, 2019, pp. 271-279. (in Russ.).

7. Baranov V. A., Dubovtsev S. V. *Modul' statistiki informatsionno-analiticheskoi sistemy "Manuskript": funktsii i demonstratsiya dannykh* [Statistics module of the information-analytical system “Manuscript”: functions and data demonstration]. *Informatsionnye tekhnologii i pis'mennoe nasledie: materialy IV mezhdunar. nauch. konf.* (Petrozavodsk, 3–8 sentyabrya 2012 g.) / otv. red. V. A. Baranov, A. G. Varfolomeev [Information technology and written heritage: materials of the IV international scientific conference (Petrozavodsk, September 3–8, 2012) (eds. Baranov V. A., Varfolomeev A. G.]. Petrozavodsk; Izhevsk, 2012, pp. 23-26. (in Russ.).

8. *Svodnyi katalog slavyano-russkikh rukopisnykh knig, khranyashchikhsya v SSSR (XI–XIII vv.)* [The consolidated catalog of Slavic-Russian manuscript books stored in the USSR (11th–13th cent.)]. Moscow, Nauka Publ., 1984, 406 p. (in Russ.).

9. Baranov V. A., Zholobov O. F. [Linguistic and statistical study of frequency words in the Words of Kirill Turovsky (according to the manuscript of the National Library of Russia, F.p. I. 39)]. *Slověne = Slov□ne*. International Journal of Slavic Studies. In the press. (in Russ.).

10. Zholobov O. F. *O kontrastiruyushchikh orfograficheskikh sistemakh v rukopisi XIII v. (k internet-izdaniyu Tolstovskogo sbornika)* [On contrasting spelling systems in the manuscript of the 13th century (to the online edition of the Tolstoy miscellany)]. *Drevnyaya Rus'. Voprosy medievistiki* [Ancient Russia. Questions of Medieval Studies]. 2018, 3 (73), Pp. 77–89. (in Russ.).

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

12. Paul A. and Jr. Gore. *Cluster analysis*. In: Handbook of Applied Multivariate Statistics and Mathematical Modeling. (Eds.) Howard E.A. Tinsley and Steven D. Brown. Academic Press, 2000. Pp. 297-321. (in Eng.).

13. Tryon, R. *Cluster analysis*. New York: McGraw Hill, 1939. (in Eng.).

14. Cattell, R. B. *A note on correlation clusters and cluster search methods*. Psychometrica, 9, 1944. Pp. 169-184. (in Eng.).

15. Sokal, R. and P. Sneath. *Principles ofnumeric taxonomy*. San Francisco: W. H. Freeman, 1963. (in Eng.).

16. Golovin B. N. Op. cit. Pp. 159–166.

17. Uspenskii B. A. *Istoriya russkogo literaturnogo yazyka (XI–XVII vv.)* [History of Russian literary language (11th–17th cent.)]. Budapest : Tankönyvkiadó, 1988, 451 p. Pp. 18, 68. (in Russ.).

18. Picchio, R. *Models and patterns in the literary tradition of Medieval Orthodox Slavdom*. In: American contributions to the Seventh International Congress of Slavists, II. The Hague, 1973. P. 445. (in Eng.).

19. Pichkhadze A. A. *Perevodcheskaya deyatel'nost' v domongol'skoi Rusi: lingvisticheskii aspekt* [Translation Activities in Pre-Mongol Rus: Linguistic Aspect]. Moscow : NP «Rukopisnye pamyatniki Drevnei Rusi», 2011, 408 p. P. 54. (in Russ.).

* * *

The Oldest Russian Manuscripts as an Object of Statistical Analysis

V. A. Baranov, DSc in Philology, Professor, Kalashnikov ISTU, Izhevsk, Russia
O. F. Zholobov, DSc in Philology, Professor, Kazan Federal University, Kazan, Russia

The work describes two statistical experiments aimed at revelation of the correlation proximity/distance of 12 texts, survived in the Russian copies of the 11th century, and their comparison with the works of Kirill Turovsky – the author of the 12th century – (RNB, F.p.I. 39, 13th cent.; ff. 1–48). The paper presents the results of the comparative analysis of a) various ways of extraction of linguistic units from texts and b) retrievals of various volumes and also of the linguistic interpretation of basic laws of manuscript grouping.

The degree of the statistic-linguistic neighboring of the manuscripts is computed in two stages: at the first stage the lists of the most frequent words of each pair of texts are compared (computation of Spearman's rank correlation coefficient), at the second stage the texts are grouped on the basis of the obtained correlation values which are taken as distances between the manuscripts (cluster analysis is applied and a dendrogram is plotted).

The extraction of the most frequent words of the manuscripts, the development of ranked lists, obtaining the data on the quantity (and the rank, respectively) of each of the forms in other codices are carried out by means of the statistics module of the historical corpus "Manuscript". Computation of the correlation coefficients of the texts and clustering texts are done with the help of software package "Statistics" (TIBCO Software Inc.). Lists of various volumes (from 50 to 300 word forms) and comprising units of various degrees of unification relative to the text forms were analyzed.

The result of the first experiment was the revelation of three main stable clusters of the sub-corpus: the group of Gospels, the group of Menaia and the group of miscellanies of various contents.

The second experiment gave a possibility of seeing the dependence of the proximity of the sermons of Kirill Turovsky to various clusters on the degree of unification of forms in the retrievals and the retrievals volumes.

The linguistic analysis of the results was a basis for revelation of lexical-grammatical and lexical-semantic factors determining occurrence of the texts of Kirill Turovsky in different clusters at various initial conditions of retrieval: in the group of Gospel copies (at the retrieval volume from of 50 or 100 words), in the sub-group of miscellanies (at the retrieval of 200 words), in the sub-group of Izbornik 1073 and The Pandects of Antiochus (the retrieval of 300 words).

Keywords: linguistic statistics, ancient Russian texts, XIth century, Kirill Turovsky.

Получено: 29.10.19