Применение квантитативных корпусных методик для выявления церковнославянизмов в современном русском языке[1]

Литвинцева К. В. (tinalitvina@gmail.com)
НИУ ВШЭ, Москва, Россия
Ляшевская О. Н. (olesar@yandex.com)
НИУ ВШЭ, Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Use of quantitative corpus methods for detection of Slavonicisms in modern Russian

Litvintseva Kristina (tinalitvina@gmail.com)
National Research University Higher School of Economics, Moscow, Russia
Lyashevskaya Olga (olesar@yandex.com)
National Research University Higher School of Economics,
Vinogradov Institute of the Russian Language RAS, Moscow, Russia

В основе данного исследования лежит гипотеза о дискурсивной близости церковнославянского языка и религиозного христианского дискурса современного русского языка. Будет сделана попытка при помощи корпусного статистического анализа показать, что, с точки зрения лексического состава, эта часть языка заметно сближается с церковнославянским языком, если сравнивать ее с неспециализированным современным русским языком. Это может служить доказательством специфичности исследуемой части языка, дополнительным доводом при решении вопроса о его отдельном статусе. Исследование проводится на материале Национального корпуса русского языка, а именно через сравнение данных Церковнославянского корпуса, Основного корпуса и входящего в него подкорпуса церковно-богословских текстов. Сопоставляется употребимость часто встречающихся лексем церковнославянского языка с частотами их лексических аналогов в современном русском языке в целом, а также в заданном подкорпусе церковно-богословских текстов, представляющих религиозный христианский дискурс современного русского языка. С помощью критерия логического правдоподобия и метода

[1] Исследование выполнено при финансовой поддержке РГНФ, грант № 17-04-12064 «Разработка модулей НКРЯ для автоматической разметки и словарной поддержки старорусских и церковнославянских текстов».

главных компонент (PCA) выявляется пласт лексики современных текстов, которые предлагается считать церковнославянизмами.

Ключевые слова: корпусные исследования, квантитативные корпусные методы, значимая лексика, церковнославянский язык, современный русский язык, религиозный христианский дискурс

The starting point of the study is the hypothesis of a discursive proximity of Church Slavonic and Christian religious discourse of the modern Russian language. An attempt is made by means of quantitative corpus analysis to show that, from the point of view of lexical structure of this part of the language is closer to Church Slavonic than the mainstream modern Russian language. This can serve as a proof of the specificity of the register in question, an additional argument when deciding on its separate status. Research is based on the material of the of the Russian National Corpus, namely, the Church-Slavonic corpus, the Main corpus and the subcorpus of church-and-theology texts. The frequency of most frequent tokens of the Church Slavonic language is compared against the occurency of their lexical analogues in the modern Russian language, presented in the Main corpus in the given Subcorpus of church-theological texts, that is, in the religious discourse of the modern Russian language. Using the log-likelihood criterion and PCA visualizations, we reveal the body of lexemes in Russian texts that can be considered Slavonicisms (tserkovnoslavyanizmy).

Key words: corpus study, quantitative corpus methods, lexical markers of discourse, Church Slavonic language, modern Russian language, religious discourse

1. Введение

Данное исследование посвящено изучению лексики религиозного христианского дискурса современного русского языка (XVIII–XXI вв.) – то есть языка проповедей, богословских эссе, религиозной прессы и других текстов, созданных христианами для обсуждения христианства – как отдельного пласта современного русского языка. Будет сделана попытка при помощи корпусного статистического анализа показать, что, с точки зрения лексического состава, эта часть языка заметно ближе к церковнославянскому, чем к неспециализированному современному русскому языку, что может служить доказательством специфичности исследуемой части языка, дополнительным доводом при решении вопроса о его отдельном статусе. В основе данного исследования лежит гипотеза

о дискурсивной близости церковнославянского языка и религиозного дискурса современного русского языка.

Необходимость подобных статистических сравнительных исследований обсуждалась И. С. Улухановым в отношении древнерусских текстов: «Вопрос о том, какие из элементов церковнославянского языка легче выходят за его пределы, остается недостаточно изученным. На ограниченном материале была выявлена достаточно очевидная закономерность использования славянизмов в древнерусском языке, которая, однако, не могла быть сформулирована без статистических данных об употребительности церковнославянской лексики: чем чаще слово употреблялось в памятниках церковнославянского языка, тем дальше оно проникает за их пределы – в летописные рассказы, в деловые памятники, в устную речь <...> В дальнейшем необходимо фронтальное изучение устоявшихся особенностей употребления различных единиц церковнославянского языка, детальный статистический анализ этих единиц в их соотношении с древнерусскими единицами. Это откроет перспективы обоснованного, конкретного и объективного решения вопроса о роли славянизмов в истории русского языка и о системе разновидностей языка Древней Руси».[2] Ср. также с мнением В.В. Леденевой, высказанным ею относительно авторского идиостиля, однако вполне применимым к специфическим стилистическим дискурсивным исследованиям: «Предпочтение языковых средств, устанавливаемое по фактору частотности, представляет собой особенности идиостиля. Заметное место занимает стилистически окрашенная лексика, призванная усилить или передать то или иное впечатление, оценку, характеристику, художественно-эстетический эффект».[3]

2. Материал исследования

При работе в системе Национального корпуса русского языка (НКРЯ) есть возможность задать пользовательский подкорпус с необходимыми параметрами. Если в фокусе исследования находится религиозный дискурс современного русского языка, то наиболее оправданным будет обращение именно к подкорпусу с параметром «церковно-богословская сфера функционирования» (далее ЦБ). В этот подкорпус

---

[2] *Улуханов И.С.* Церковнославянский язык русской редакции: сфера распространения и причины эволюции // Исследования по славянским языкам. 2003. № 8. С. 17, 20.

[3] *Леденева В.В.* Идиостиль как система отношений // Вестник Тамбовского университета. Серия: Гуманитарные науки. 2001. № 5. Т. 23. С. 12.

группируются тексты религиозной тематики: обычно речь в них идет о Боге, догматах церкви, богослужении, церковной жизни и под. По этой причине в сгруппированных в ЦБ-подкорпус текстах достаточно легко вычленимы лексические маркеры – слова, называющие соответствующие реалии (*Бог, церковь, потир, батюшка*). Среди таких лексических маркеров особое место занимают церковнославянизмы – лексемы, пришедшие из богослужебной практики и используемые в неиронической, нестилизаторской функции, но в рамках дискурсивной коммуникации.

Тексты, группируемые в ЦБ-подкорпус, не однородны по составу и времени написания. Так, высокая активность в создании текстов церковно-богословской сферы связана с деятельностью митр. Платона (Левшина), XVIII в., прот. С. Н. Булгакова, прот. Георгия Флоровского, митр. Антония (Сурожского), XX в., с публикацией Синодального перевода Евангелия в XIX в. и изданием Деяний Священного Собора в начале XX в. Среди жанров, представленных в ЦБ-подкорпусе НКРЯ, преобладает проповедь, однако также имеются тексты таких жанров, как Священное писание, послание, поучение, слово, житие, статья, трактат, дневник, указатель и др. В целом такое распределение призвано отражать реальную языковую картину, в соответствием с принципами сбалансированности НКРЯ[4].

Вышеперечисленные особенности среди прочего указывают на необходимость при анализе языковых особенностей текстов церковно-богословской сферы учитывать некоторое влияние языка XVIII в. и его особенностей, не регистрируемых в современном языке, таких, как: большое количество элементов фонетического письма, большое количество устаревшей лексики и под.

Следует добавить, что только один из 942 текстов, отнесенных к религиозной сфере функционирования, посвящен изучению ислама, однако он не оказывает существенного влияния на картину распределения лексики в религиозном дискурсе. Кроме того, как показала Ж.К. Киынова, в переводах исламских религиозных текстов на русский язык славянизмы используются так же, как и в христианских текстах[5], т. е., несмотря на некоторое превалирование собственно исламских терминов, картина

---

[4] Ср. «корпус содержит по возможности все типы письменных и устных текстов, представленные в данном языке <…> все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода». См. *Что такое Корпус?* / Национальный корпус русского языка. [Электронный ресурс] URL: http://www.ruscorpora.ru/corpora-intro.html

[5] *Киынова Ж.К.* Славянизмы как средство стилизации в переводах религиозной литературы // Вестник Московского университета. Сер. 22. Теория перевода. 2014. № 1. С. 62–69.

распределения лексики в исламских церковно-богословских текстах и в христианских церковно-богословских текстах существенно не различается.

Все особенности церковно-богословских текстов ЦБ-подкорпуса, перечисленные выше, имеют локальный характер и не должны оказать существенного влияния на общую картину представления лексики в религиозном дискурсе современного русского языка. В то же время в основе данного исследования лежит гипотеза об особом влиянии церковнославянского языка на религиозный дискурс современного русского языка (а не только на тексты, порожденные в XVIII в.). Иначе говоря, в текстах, составляющих религиозный дискурс представлено сравнительно большое количество церковнославянизмов, употребляемых в неиронической и нестилизаторской функции как при цитировании, так и в собственной речи.

Сопоставим примеры употребления церковнославянизмов в современном русском языке.

В примере (1) представлен отрывок из газетной статьи, тема которой определяется НКРЯ как ‘администрация и управление’, таким образом, ее нельзя отнести к религиозному дискурсу. Церковнославянское по происхождению «аки агнцы» стилистически вписывается в иронический контекст, соседствуя с употребленной в переносном значении лексемой *правильные*, стилизацией разговорной речи *они ж у нас; не к ночи будь помянуты; в общем, пишите; а потом сочтемся*. Общая стилистика отрывка ироническая. Приведенный пример употребления славянизма[6] в современном тексте на русском языке – классический случай иронической функции[7].

(1) *Предлагайте, возражайте, ругайте. Но в меру, чтоб депутаты не огорчались. Они ж у нас правильные: ни мат, ни табак, ни дом за границей — ничего не признают.* ***Аки агнцы****, не к ночи будь помянуты. В общем, пишите. А потом сочтёмся…* [Геннадий Рявкин. Это странное ремесло (2013.04.05) // «Новгородские ведомости», 2013, ОК НКРЯ]

Известно, что в церковнославянском языке *агнец* используется в прямых значениях: ‘ягненок’, ‘жертва’, ‘Иисус Христос’.[8] В религиозном дискурсе, реализуемом в

---

[6] Церковнославянизм мы расцениваем как частный случай славянизма. См. об этом, например, *Литвинцева К.В.* «Церковнославянизм» как лингвистический термин // Вестник Орловского государственного университета. 2015. № 6 (47). С. 264–267.

[7] Об иронической функции славянизмов см. в работах: *Замкова В.В.* Славянизм как стилистическая категория в русском литературном языке XVIII в. Л., 1975; *Семенов П.А.* Проблема классификации стилистических функций славянизмов (диахронический аспект) // Вестник Новгородского гос. ун-та. Серия: Гуманитарные науки. 1998. № 4. С. 134–138.

[8] См.: *Дьяченко Г., прот.* Полный церковнославянский словарь с внесением в него важнейших древнерусских слов и выражений. М., 1993 (репринт). С. 5; *Поляков А.Е.* Грамматический словарь

церковно-богословских текстах, лексема *агнец* используется либо в цитатах церковнославянского происхождения (2), либо в тех же значениях, что и в церковнославянском языке[9]. Так, в примерах (3) и (4) лексема *агнец* используется по отношению к Иисусу Христу.

(2) *И значительны, знаменательны слова, которые провозглашает Иоанн Креститель: «Вот* ***Агнец*** *Божий, Который берет на Себя весь грех мира, Который на Свои плечи берет мир с его грехом, со всеми последствиями этого греха»…* [митрополит Антоний (Блум). Крестный путь Христов (1992), ОК НКРЯ]

(3) *Как* ***Агнец*** *излиял Свою Кровь вместо* ***агнцев****, закалаемых в пустыне для жертвы, и принес Собою жертву Богу Отцу за спасение всего мира; как человек был положен во гробе, а как Бог освятил олтарь Церкви из язычников; как царь был охраняем стражами и запечатленный лежал во гробе, но как Бог чрез ангельские воинства вещал бесовским силам втвердыне ада: «Возмите врата князи ваша, и возмитеся врата вечная: и внидет Царь славы»* [Слово на Святую Пасху (2004) // «Журнал Московской патриархии», 2004.04.26, ОК НКРЯ]

(4) *Стра́нно и вели́ко неслóбіе показа̀, пребл҃же́нный, єгда̀ єди́нъ бу́ій и слонра́вный уда́ри єго̀ въ лани́ту: о́нъ же,* ***а́гнцу*** *хрⷭ̂ту̀ поревнова́въ, до землѝ поклони́ся бію́щему, моля̀ бг҃а ѡ прощє́нiи є́му̀.* [Акафист святому Тихону Воронежскому, ЦК НКРЯ]

Таким образом, в религиозном дискурсе современного русского языка церковнославянская по происхождению лексика используется в тех же значениях, что и в церковнославянском языке или в приближенных к ним[10].

3. Методы исследования

3.1. Статистические методы исследования

Исследование базируется на понятии “значимости лексики”. Для определенного сегмента языка возможно выделить определенные лексические маркеры: как отмечено в

церковнославянского языка (по материалам корпуса). [Электронный ресурс] URL: http://dic.feb-web.ru/slavonic/dicgram/

[9] См.: *Добрушина Е.Р., Литвинцева К.В., Польсков К.О., Хангиреев М.А.* От «аббата» до «аналоя»: фрагмент лингво-энциклопедического словаря русской христианской лексики // Вестник ПСТГУ. Серия III. Филология. 2011. № 3 (25). С.119–148.

[10] См. об этом в работах: *Добрушина Е.Р.* Словарь христианской лексики: состав словника // Вестник ПСТГУ. Серия III: Филология. 2012. № 3(29). С. 105–113; *Литвинцева К.В.* Особенности функционирования трех фразеологизмов с лексемой *Божий* в религиозных и светских текстах // Вестник ПСТГУ. Серия III: Филология. 2014. № 4 (39). С. 67–81.

Предисловии к Новому частотному словарю русской лексики «частота слов *процесс* и *теория* в научных публикациях значительно превышает их частоту во всех остальных текстах корпуса. Аналогичным образом, слова *ну, да, вот, пожалуйста* употребляются чаще в устной речи, а слова *сказать, спросить, локоть, снег* – в художественной литературе».[11] Полученные списки таких лексем называют значимой лексикой.

Сам по себе факт частоты той или иной лексики в том или ином сегменте языка еще не говорит об особой конституирующей роли данной лексики для данного сегмента либо об уникальности самого сегмента, потому что наиболее частотные служебные слова приблизительно равномерно употребляются в текстах разных стилей и жанров. Однако сравнивая частоты слов в разных подкорпусах, можно получить списки значимой лексики для того или иного сегмента языка, которые покажут реальную картину распределения слов-маркеров в анализируемом сегменте.

В нашем исследовании объектами сравнения выступают Церковнославянский корпус НКРЯ (далее ЦК) и ЦБ-подкорпус. Церковнославянский корпус НКРЯ – самый объемный из его Исторических корпусов, объем на момент исследования (август 2017) – 4,7 млн. словоупотреблений. Объем ЦБ-подкорпуса на момент исследования – 4 млн. словоупотреблений. В качестве референтного корпуса используются тексты Основной корпус НКРЯ (далее ОК), созданные в 1950-2007 гг. Данные по этому корпусу получены из словаря Ляшевская, Шаров 2009 (объем выборки – 92 млн. словоупотреблений).

Статистическое сравнение частотности лексики по этим корпусам представляется информативным с точки зрения выявления общих полей. Что особенно ценно, именно ЦК можно охарактеризовать как аутентичный корпус, т.к. именно кодифицированные тексты, созданные в разные периоды на церковнославянском языке, возможно собирать в корпус, покрывающий очень существенную часть всех функционирующих на этом языке текстов, то есть обладающей реальной репрезентативностью, в отличие от, например, ОК НКРЯ. Составители ОК при выборе объемов помещаемых в него текстов определенных типов с определенной датировкой стремятся к сбалансированности, моделированию состояния реального языка, но соответствующие методики пока мало разработаны, а электронные текстовые ресурсы в отдельных случаях труднодоступны, поэтому ОК, являясь, конечно же, значимым свидетельством о различных свойствах современного русского языка, все же очень далек от того, чтобы представить современный русский язык в полной мере. Что

[11] *Ляшевская О.Н., Шаров С.А.* Введение к частотному словарю современного русского языка. М., 2009. С. viii.

же касается ЦБ, то он достаточно объемен и при этом однороден по составу, поэтому в большей мере, чем ОК, хоть и в меньшей, чем ЦК может считаться адекватно представляющим соответствующий ему срез современного языка.

В качестве метрики сравнения используется критерий отношения правдоподобия (log–likelihood), вычисляемый на основе следующей матрицы:

	Подкорпус 1	Подкорпус 2
Частота	*a*	*b*
Размер	*c*	*d*

Таблица 1. Матрица абсолютных частот для вычисления коэффициента значимости LL-score.

Эта метрика, широко используемая в корпусной лингвистике[12], опирается на математическое ожидание частоты слова, исходя из доли вхождений слова в совокупном корпусе и относительного размера рассматриваемого подкорпуса. Например, ожидаемая частота E1 слова *се́рдце* в ЦБ составляет 1186 словоупотреблений (3967298 · 28677 / 95949705), а наблюдаемая частота – 6132 словоупотреблений, т. е. более чем в 5 раз больше. Напротив, ожидаемая частота E2 того же существительного в ОК составляет 27491 словоупотреблений, что больше наблюдаемой частоты в этом корпусе (22545). Показатель критерия log–likelihood (LL-score) для этого слова составляет 11207.95, что значительно выше порога статистической значимости. При расчете метрики мы не принимаем в расчет, что корпуса ЦБ и ОК частично пересекаются, поскольку вклад ЦБ в ОК пренебрежимо мал.

	ЦБ	ОК	Всего
Частота	6 132	22 545	28 677
Размер	3 967 289	91 982 416	95 949 705

Таблица 2. Частота существительного *се́рдце* в ЦБ и ОК и размер соответствующих корпусов.

[12] *LL-score* = $2 \cdot (a \cdot \ln(a / E1) + b \cdot \ln(b / E2))$; где $E1 = c \cdot (a+b)/(c+d)$; $E2 = d \cdot (a+b)/(c+d)$.
См. *Rayson P., Garside R.* Comparing corpora using frequency profiling // Proceedings of the Comparing Corpora Workshop at ACL 2000. Hong Kong, 2000. P. 1–6; *Ляшевская О.Н., Шаров С.А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. Отношение правдоподобия учитывает как относительную частоту (во сколько раз чаще слово встречается в одном корпусе по сравнению с другим), так и абсолютную частоту. Как отмечается в предисловии к частотному словарю НКРЯ, «[п]оследнее обстоятельство важно, поскольку значимость того, что слово встретилось в подкорпусе в 10 раз чаще чем в основном корпусе, зависит от того, имеем ли 5 или 500 вхождений этого слова в подкорпус. В первом случае она может быть связана со случайными флуктуациями, во втором эти данные статистически значимы. Достоинством критерия правдоподобия является и то, что возможна статистическая оценка значимости различия частот в подкорпусе и остальном корпусе. Если этот показатель превышает 15.31, с вероятностью более 99% можно отвергнуть гипотезу, что разница в частоте случайна и она не обусловлена существенными различиями в составе корпуса» (Там же. *viii*).

В качестве еще одного способа анализа ассоциации лексики с тем или иным из рассматриваемых корпусов был выбран метод главных компонент (principal component analysis, PCA)[13]. Метод основан на понятии расстояния (χ^2) в векторном пространстве, в котором каждое исходное измерение задается частотами слов в определенном корпусе. В нашем случае, три измерения задаются осями ЦС, ОК и ЦБ. Чем меньше угол между вектором слова и одной из осей, тем больше слово ассоциируется с соответствующим корпусом, иными словами, его вклад в корпус. Метод PCA позволяет перевести координаты точек-лексем на плоскость таким образом, что можно визуально выделить кластеры, ассоциированные с тем или иным корпусом. Перед применением метода абсолютная частота слов была логарифмирована.

3.2. Установление лексических соответствий

В качестве предварительной процедуры словники корпусов были лемматизированы, а именно, при помощи морфологического анализатора Mystem[14] все словоформы одного слова были приведены к начальной, словарной форме (лемме). Ошибки лемматизации были устранены вручную, кроме того, вручную были разведены леммы глаголов разного вида.

Затем были найдены соответствия лемм в трех корпусах. Под лексическими соответствиями (аналогами) мы понимаем условно идентичные по внешней и внутренней форме лексемы, морфологическая форма которых одинакова, а основные значения не имеют существенных отличий (*агнецъ—агнец, слово—слово, пріимати—принимать*). Здесь следует сделать несколько оговорок. Некоторые из сопоставляемых здесь лексем могут быть признаны вслед за О.А. Седаковой церковнославяно-русскими паронимами, т. е. близкими по написанию и звучанию словами родственных языков[15] (например, соотношение подзначений в отношении и набора, и частот использования у церковнославянских лексем *сын* и *слово* принципиально отличается от соответствующего соотношения в современном русском языке. Мы также отдаем себе отчет в том, что лексические единицы типа *слава* в большинстве случаев имеют разную семантику в

---

[13] См. Levshina N. How to do Linguistics with R. John Benjamins, 2015. P. 353-361. Для анализа данных и построения графиков использовался язык R, библиотеки FactoMineR и factoextra.

[14] См.: Mystem+ [Электронный ресурс] URL: http://web-corpora.net/wsgi/mystemplus.wsgi/mystemplus/.

[15] *Седакова О.А.* Предисловие / Словарь трудных слов из богослужения: Церковнославяно-русские паронимы. М., 2008. С. 1-12.

церковнославянском и современном русском языках, однако они также являются частью исследования как имеющие определенно сходную семантику в церковнославянском языке и религиозном дискурсе современного русского языка (отражаемом в ЦБ-подкорпусе). Впрочем, в большинстве случаев семантическая разница между лексическими аналогами в церковнославянском и русском языках не оказывает существенного влияния на общую картину распределения лексики. По этой причине в данном исследовании используется термин «лексический аналог» для обозначения внешне схожих лексических единиц близкой семантики.

В ходе построения частотного списка ЦБ-подкорпуса нами были приняты некоторые технические решения, призванные сделать его единообразным. Среди таких решений следует назвать:

1. Нивелирование паронимов. В отличие от словаря О.А. Седаковой перед нами не стояли просветительские задачи толкования значения, нам необходимо было получить общую картину представления церковнославянизмов в религиозном дискурсе современного русского языка. Поэтому, например, русским лексическим аналогом церковнославянской лексемы *добрый*, часто употребляемой в значении ‘красивый’, мы считали лексему *добрый*, обычно употребляемую в значении ‘хороший’.

2. Унифицирование морфемного состава лексических аналогов. Исходя из известных фактов о морфемном составе церковнославянизмов, таких, как: неполногласие; начальные *а-, у-,* э-; *щ* вместо русского *ч*; *жд* вместо *ж*; специфические окончания типа *–ие* и др.[16], мы приняли решение выровнять русские лексические аналоги по принципу отсутствия таких признаков у лексем. Так, например, церковнославянской лемме *премѣнѧти* в нашей системе соответствует лемма *переменять*.

Естественно, что такая унификация сопряжена с возникновением некоторых проблем и «шероховатостей». Например, при наличии двух аналогов, как в случае с церковнославянской леммой *сребряный*, для которой в русском языке сосуществует два аналога: собственно церковнославянизм *сребряный* и собственно русский вариант – *серебряный*, нам необходимо принять решение в пользу одной леммы. У такого решения имеется техническая причина возникновения избыточности при подсчете: наличие двух аналогов одной леммы вдвое увеличит ее шансы на попадание в топ значимой лексики и, соответственно, исказит общую картину. В случае с леммой *сребряный* (как и в большинстве аналогичных) решение было принято в пользу заметно преобладающего по

[16] *Шахматов А.А.* Очерк современного русского литературного языка. М., 1941. С. 71.

частоте аналога *серебряный*. Резюмируя, следует признать, что лексическим аналогом церковнославянской леммы практически всегда становилось ее исконно русское однокоренное соответствие, однако каждый трудный случай всегда становился предметом особого внимания.

4. Значимая лексика церковнославянского и церковно-богословского корпусов

Мы сравнили попарно корпуса ЦС и ОК, а затем корпуса ЦБ и ОК, чтобы выявить значимую лексику ЦС и ЦБ-корпусов на фоне лексического фонда современного русского языка. В таблице 3а приведен список полнозначных слов ЦС-корпуса, отсортированный в порядке убывания LL-score, в таблице 3б – аналогичный список для ЦБ корпуса (здесь и далее леммы приводятся в современном написании).

Табл. 3. Значимая лексика ЦС и ЦБ корпусов, упорядоченная по убыванию коэффициента LL-score.

(а) Значимая лексика ЦС-корпуса

Лемма	**ipm, ЦС**	**ipm, ОК**	**#LL-score**
господь	8 594	58	211 866
глас	25 387	6	148 526
бог	40 265	425	137 359
святой	27 230	140	115 709
ныне	21 401	56	103 930
слава	23 996	120	102 519
глаголать	14 294	1	85 524
христос	19 970	101	85 081
ирмос	11 381	1	68 233
божественный	13 096	36	63 055
радоваться	12 152	58	52 550
дева	9 785	16	50 738
богородица	8 854	8	48 241
приять	8 119	3	46 700
молитва	9 691	52	40 735

(б) Значимая лексика ЦБ-корпуса

Лемма	**ipm, ЦБ**	**ipm, ОК**	**#LL-score**
бог	6 133	425	73 846
божий	3 437	100	56 887
христос	2 906	101	45 587
господь	2 451	58	42 834
церковь	3 097	179	40 470
святой	2 264	140	28 750
иисус	1 496	34	26 432
молитва	1 508	52	23 737
дух	1 925	154	21 652
грех	1 266	72	16 692
вера	1 645	165	16 316
духовный	1 382	104	16 029
христов	913	26	15 256
апостол	832	22	14 160
преподобный	807	26	13 021

вопиять	6 767	2	39 585
душа	16 186	357	38 981
петь	11 698	143	37 668
божий	10 482	100	37 122
преподобный	7 920	26	37 069
господень	6 179	1	36 744
тропарь	5 976	3	33 736
сын	13 510	285	33 374
единый	11 082	161	33 113
стих	6 352	11	32 551

человек	6 281	2 723	12 674
евангелие	784	25	12 582
жизнь	4 054	1 390	12 477
учение	861	47	11 542
любовь	1 758	324	11 237
сердце	1 546	245	11 208
благодать	610	13	10 970
душа	1 716	357	9 792
христианин	633	23	9 754
старец	587	17	9 731

Среди значимой лексики как ЦС-корпуса, так и ЦБ-подкорпуса мы наблюдаем слова семантически соотносимые с Богом (*бог, божий, господь*), церковью (*молитва*), святыми (*святой, преподобный*), человеком (*душа*). Однако мы не можем говорить об абсолютном совпадении лексики в анализируемых корпусах, причиной чему в большей мере служит факт попадания в верхнюю часть списка значимой лексики ЦС-корпуса собственно богослужебной лексики (*глас, ныне, слава, ирмос, радоваться, петь, тропарь, стих*). Такое распределение значимой лексики в корпусах объясняется тем, что церковнославянский язык - это в прежде всего богослужебный язык, тогда как в современном русском языке религиозный дискурс служит скорее для комментирования богослужения, Священного Писания, догматов и т. п. В то же время в целом на основании сопоставления значимой лексики ЦС-корпуса и ЦБ-подкорпуса мы можем говорить не только о лексических аналогах в церковнославянском языке и религиозном дискурсе русского языка, но и о наличии общих семантических полей, включающих лексику близкой семантики, группирующуюся вокруг таких ядерных слов как *Бог, церковь, святой, человек*.

Следует отдельно оговорить, что лексема *приять*, несмотря на наличие лексического аналога в русском языке (*принять*) была проанализирована именно в форме церковнославянизма ввиду своей высокой частотности в ОК. А для лексемы *вопиять*, например, не учитывался формальный аналог *вопить* по причине расхождения в семантике (взывать vs. орать).

Среди служебной лексики значимыми оказываются, во-первых, замещенные в современном русском языке другими словами местоимения (*аз, сий, сие, той, иже*), союзы (*яко*, *ибо, аще*), неполногласные предлоги (*пред*, *чрез*). Во-вторых, в круг значимой лексики попадают личные и притяжательные местоимения 1-го и 2-го лица (*мы*, *наш*, *твой*), отражающие диалогические отношения между богом и человеком.

Итак, на основании полученных статистических данных можно сделать вывод о лексической близости церковнославянского языка и религиозного дискурса современного русского языка, представленного в ЦБ-подкорпусе. Статистическое сравнение частотности лексики по этим корпусам показывает, что пласт ЦБ текстов современного русского языка и церковнославянские тексты имеют общее ядро, и что это ядро в религиозном дискурсе современного русского языка составляют те лексемы, которые могут быть названы церковнославянизмами.

5. Сопоставление лексики трех корпусов с помощью метода главных компонент

Критерий LL-score не позволяет сравнивать более двух корпусов; помимо этого, он используется для выявления различий в словоупотреблении (лексических маркеров), а не сходства. С помощью метода главных компонент (PCA) мы получили три графика, отражающих относительную частоту функционирования наиболее частотных имен прилагательных, существительных и глаголов сразу в трех корпусах.

На Рис. 1-3 вектора условно изображают на плоскости три домена (в ЦС, ЦБ и ОК). Чем ближе точка, символизирующая лексему, к одному из доменов по горизонтали (первая компонента) или по вертикали (вторая компонента), тем в большей степени эта лексема ассоциируется с этим доменом. На графике, изображенном на Рис. 1, имена прилагательные, расположенные справа от начала координат (*большой, самый, другой, новый*), ассоциированы с корпусом ОК и, в меньшей степени, с корпусом ЦБ; прилагательные, расположенные слева, ассоциированы с ЦС-корпусом (*честный*, *божественный*, *преподобный*). Кластер *святой* и *божий* равно близок к ЦС и ЦБ, но явно недопредставлен в ОК. Чем дальше слово отстоит от начала координат, тем более неравномерно распределение его частоты (в пользу некоторого домена). Как видно, прилагательные *большой* и *честный* демонстрируют самое неравномерное распределение, причем *большой* недопредставлен в ЦС (ср. *велии* и др. синонимы), а честный - в ЦБ.

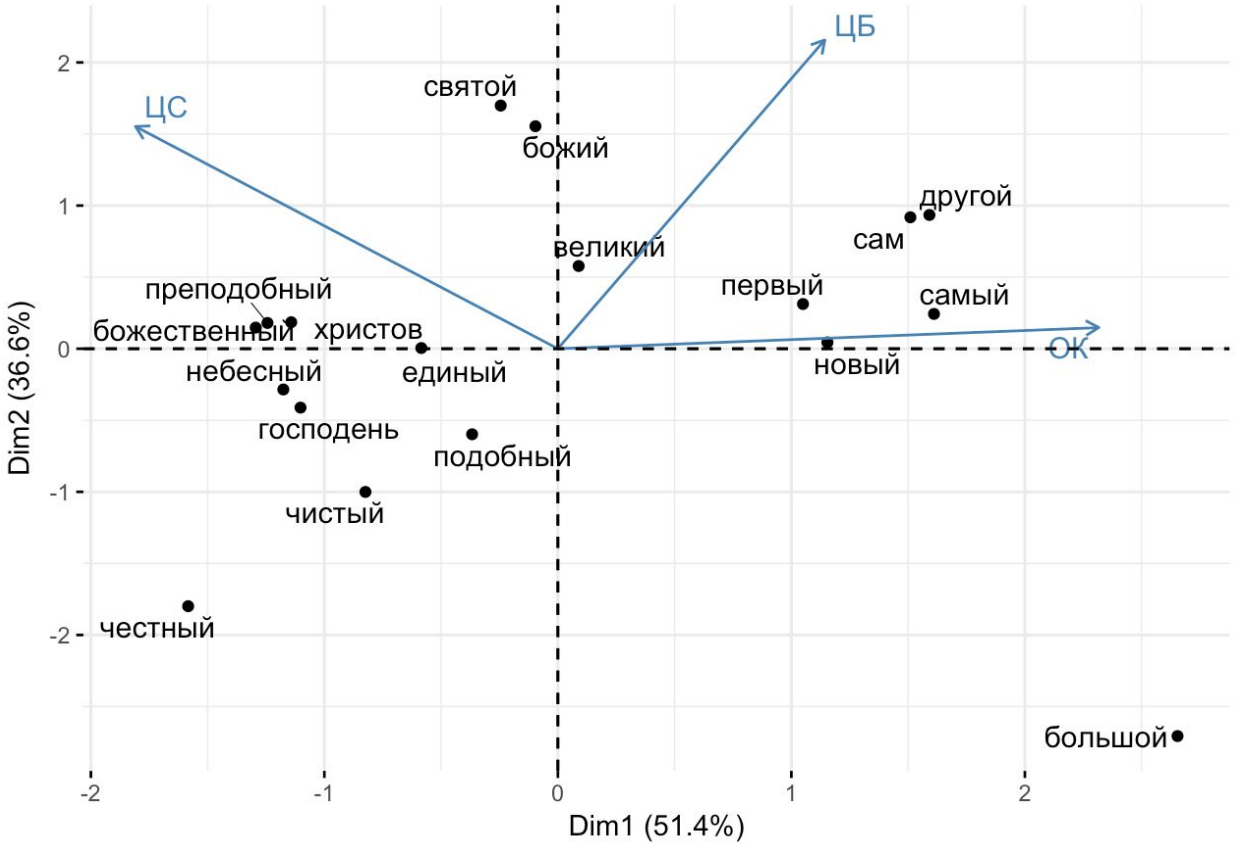

Рис. 1. Частотные имена прилагательные в корпусах ЦС, ЦБ и ОК: визуализация методом главных компонент.

Анализируя Рис. 2 аналогичным образом, можно прийти к выводу, что имена существительные *людіе*, *ирмос* и *богородичен* (т. е. богослужебная лексика, лексика акафистов, ирмологиев и т. п.) представляют в основном ЦС-корпус, но не ЦБ. Существительные *бог*, *господь* и *христос* ассоциированы и с ЦС, и с ЦБ (причем господь - больше с ЦС), а слова *раз*, *год*, *работа* “предпочитают” ОК. Вместе с тем, мы видим, что среди имен существительных наблюдается много таких, которые не обнаруживают заметных частотных преференций ни к одному из корпусов (ср. *свет*, *страсть*, *Иисус*). Это связано с тем, что, в отличие от имен прилагательных, в верхнюю часть кумулятивного списка частотной лексики попадает много высокочастотных существительных, у которых частотные распределения не зависят от рассматриваемых нами доменов.

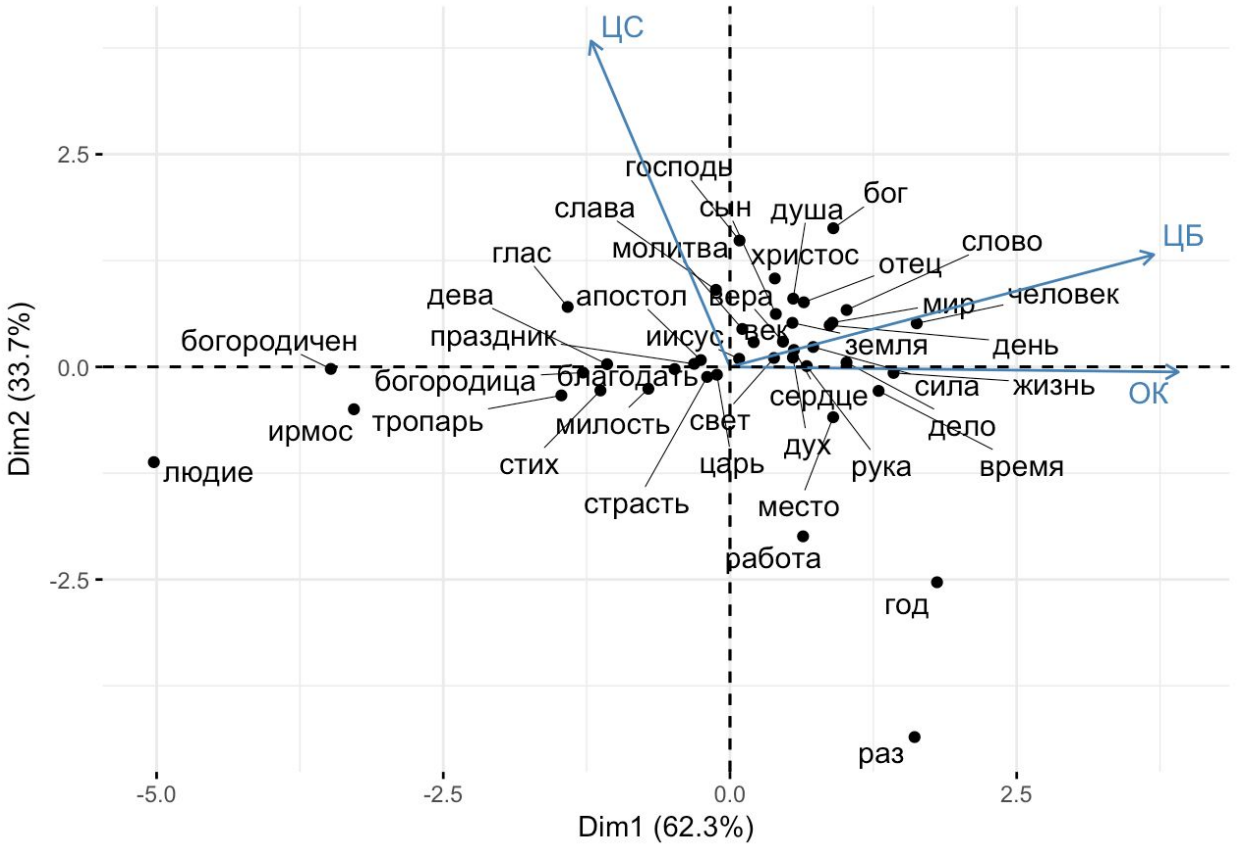

Рис. 2. Частотные имена существительные в корпусах ЦС, ЦБ и ОК: визуализация методом главных компонент.

Помимо вклада каждой из индивидуальных лексем в частоту того или иного домена, визуализация с помощью PCA позволяет оценить “расстояние” между корпусами на основе полученных частотных распределений. Вопреки первоначальной гипотезе, ЦБ-корпус неизменно оказывается ближе к ОК, чем к ЦС-корпусу, однако же его вектор неизменно отклоняется в сторону последнего. Это наблюдение можно прокомментировать таким образом, что несмотря на наличие большого пласта церковнославянизмов, в верхней части частотного списка присутствует много слов, распределение частот которых в ОК и ЦБ похоже.

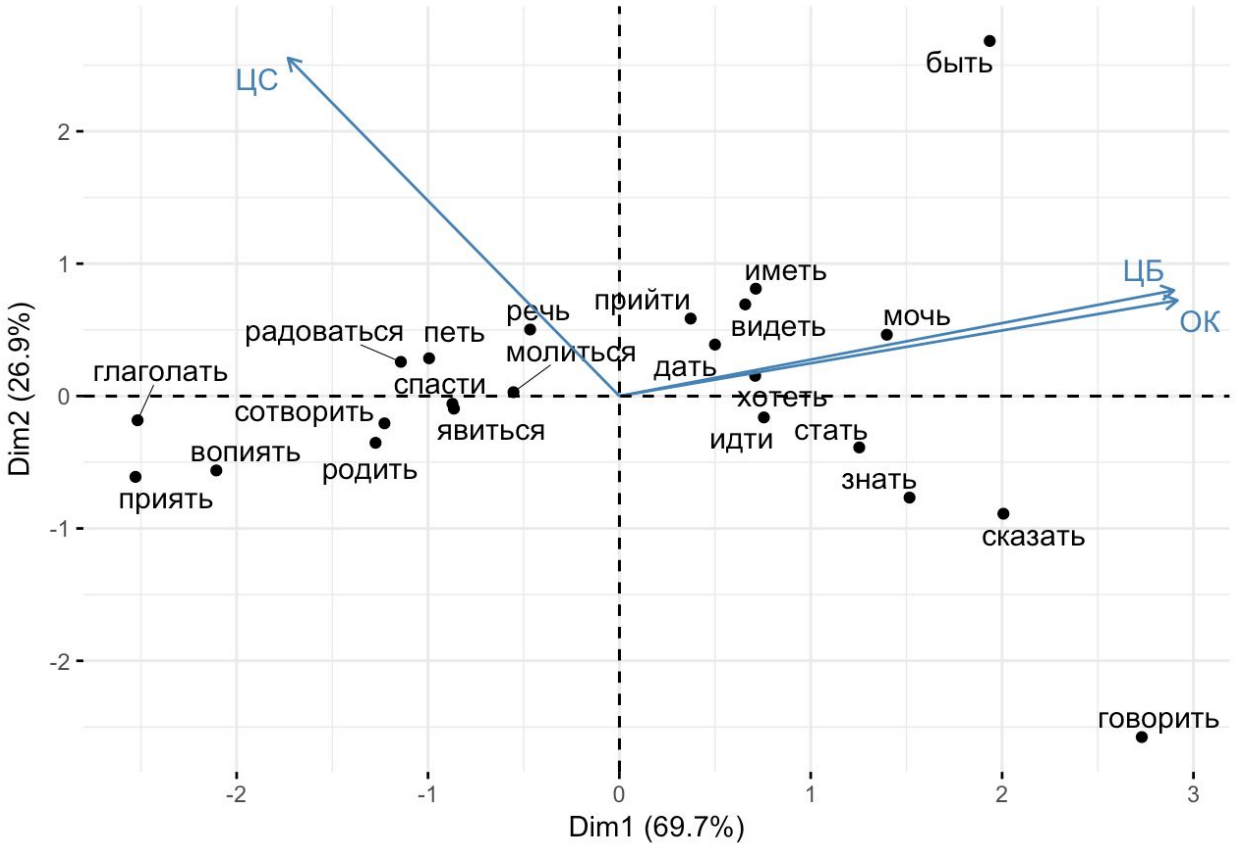

Рис. 3. Частотные глаголы в корпусах ЦС, ЦБ и ОК: визуализация методом главных компонент.

Как показывает Рис. 3, менее всего ЦБ и ОК противопоставляются по глагольной лексике. В обоих корпусах редко встречаются такие церковнославянизмы, как *глаголать*, *приять*, *вопиять*, в то время как в ЦС-корпусе реже представлены глаголы *говорить* и *сказать*. В целом, глаголы, расположенные левее начала координат, ассоциируются с ЦС-корпусом, а глаголы, расположенные справа – с ЦБ и ОК-корпусами. Расположение глагола *быть* в верхней правой четверти отражает роль этого глагола как грамматического показателя в домене ЦС, но еще больше – как конституирующего ряд более специфичных для ЦБ конструкций (ср. *человек есть образ Божий*; *этому есть причина*; модальный маркер *может быть*).

Таким образом, метод главных компонент удачно дополняет статистические методы выделения значимой лексики, позволяя оценивать “тривергенцию” (т. е. относительное распределение лексики в трех корпусах), ранжировать значимую (на фоне ОК) лексику и позволяя определить долю пересечений частотного лексического состава корпусов как корреляцию/ковариацию частотных показателей.

6. Выводы

На основе полученных данных: 1) близких по количеству совпадений списков значимой лексики для ЦБ-подкорпуса и ЦК, а также 2) сближения расстояния между этими двумя корпусами представляется обоснованным утверждать, что ЦБ-подкорпус и ЦК в достаточной мере лексически близки друг другу. Выявлена лексика русских церковно-богословских текстов, значимо отличающаяся по частоте употребления от общей лексики современного русского языка и имеющая аналоги среди лексических маркеров церковнославянского языка.

Кроме церковнославянизмов, в ЦБ-текстах был выявлен иной частотный тип лексики: теологические термины. Этот тип лексики, также являются конституирующими для ЦБ-подкорпуса и, соответственно, для религиозного дискурса современного русского языка. В то же время, это та часть значимой лексики, что выступает маркером отличия церковно-богословских текстов на русском языке от текстов на церковнославянском языке.

Некоторые вопросы, тем не менее, остаются для дальнейшего исследования. Так, в силу того, что при подсчете частотности использовались только лексемы, встречающиеся в корпусе более 10 раз, более 600 лемм остались за рамками проводимого анализа. В то же время, принимая решение об игнорировании части лексем при расчетах, мы исходили из представления о том, что если некоторая лексема частотна менее чем в 80% случаев, то вероятность ее попадания в список значимой лексики ничтожно мала.[17] Тем не менее, всегда остается небольшая вероятность того, что какая-то значимая лемма останется за границами списка, поэтому в будущем имеет смысл произвести расчеты также для низкочастотных лексем.

Еще одной проблемой примененной методики является неоднородный частеречный состав лексики. Так, в силу регулярности употребления у глаголов обычно не наблюдается значительного сдвига семантики в диахронии: глаголы типа *печь, грести, петь* сохраняют изначальное значение, тогда как существительные и прилагательные, например, достаточно часто или переходят в разряд устаревших (*очи, десный*), или изменяют семантику (*единица, добрый*). Таким образом, существительные и прилагательные,

[17] Такое представление исходит из закона Ципфа. См.: *Zipf G.K.* The Psycho–Biology of Language: An Introduction to Dynamic Philology. Boston, 1935.

являющиеся лексическими аналогами церковнославянских лексем, скорее можно будет признать церковнославянизмами, чем глаголы.

Кроме того, без последовательного контекстуального анализа всех текстов оказывается невозможным различение наречий на *–о*/*-е* (*Кате стало хорошо*) и среднего рода прилагательных на –о/-е (*платье было хорошо*), поэтому возможные морфологические различия между ними были нами проигнорированы. В дальнейшем может быть проведена работа по различению лексем данных морфологических категорий.

Безусловно, отдельного эксперимента заслуживает применение и сопоставление эффективности других статистических мер, используемых для составления контрастных ранжированных списков лексики. Ср. в этой связи, например, меры дивергенции Кулльбака-Ляйблера, энтропии Йенсена-Шеннона и др., включая варианты их оптимизации, а также методики их использования для сравнения трех корпусов (тривергенции).[18]

* * *

В соответствии с принципами открытости и воспроизводимости научных исследований данные, использованные в настоящей публикации, и скрипт R для их обработки доступны по адресу:
https://github.com/olesar/Reproducible-Research/upload/master/lexicon-of-church.

Список литературы

1. *Добрушина Е.Р., Литвинцева К.В., Польсков К.О., Хангиреев М.А.* От «аббата» до «аналоя»: фрагмент лингво-энциклопедического словаря русской христианской лексики // Вестник ПСТГУ. Серия III. Филология. 2011. № 3 (25). С.119–148.
2. *Добрушина Е.Р.* Словарь христианской лексики: состав словника // Вестник ПСТГУ. Серия III: Филология. 2012. № 3(29). С. 105–113.
3. *Дьяченко Г., прот.* Полный церковнославянский словарь с внесением в него важнейших древнерусских слов и выражений. М., 1993 (репринт).
4. *Замкова В.В.* Славянизм как стилистическая категория в русском литературном языке XVIII в. Л., 1975.

[18] См. *Oakes M. P.* Statistical Measures for Corpus Profiling // Proceedings of the Open University Workshop on Corpus Profiling, London, UK, 2008; *Mehri A., Darooneh A. H.* The role of entropy in word ranking // Physica A: Statistical Mechanics and its Applications 390(s 18–19):3157–3163, 2011; *Cabrera-Diego L. A., Torres-Moreno J.-M., Durette B.* Evaluating Multiple Summaries without Human Models: a First Experiment with a Trivergent Model. Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016.

5. *Киынова Ж.К.* Славянизмы как средство стилизации в переводах религиозной литературы // Вестник Московского университета. Сер. 22. Теория перевода. 2014. № 1. С. 62–69.
6. *Леденева В.В.* Идиостиль как система отношений // Вестник Тамбовского университета. Серия: Гуманитарные науки. № 5 Т. 23. 2001. С. 12¬¬–13.
7. *Литвинцева К.В.* «Церковнославянизм» как лингвистический термин // Вестник Орловского государственного университета. 2015. № 6 (47). С. 264–267.
8. *Литвинцева К.В.* Особенности функционирования трех фразеологизмов с лексемой Божий в религиозных и светских текстах // Вестник ПСТГУ. Серия III: Филология. 2014. № 4 (39). С. 67–81.
9. *Ляшевская О.Н., Шаров С.А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М., 2009.
10. *Ляшевская О.Н., Шаров С.А.* Новый частотный словарь русской лексики [Онлайн-версия словаря Ляшевская, Шаров 2009]. [Электронный ресурс] URL: http://dict.ruslang.ru/freq.php?
11. *Поляков А.Е.* Грамматический словарь церковнославянского языка (по материалам корпуса). [Электронный ресурс] http://feb-web.ru/febupd/slavonic/dicgram/
12. *Седакова О.А.* Словарь трудных слов из богослужения: Церковнославяно-русские паронимы. М., 2008.
13. *Семенов П.А.* Проблема классификации стилистических функций славянизмов (диахронический аспект) // Вестник Новгородского гос. ун-та. Серия: Гуманитарные науки. 1998. № 4. С. 134–138.
14. *Улуханов И.С.* Церковнославянский язык русской редакции: сфера распространения и причины эволюции // Исследования по славянским языкам. № 8. 2003. С. 1–26.
15. Что такое Корпус? / Национальный корпус русского языка. [Электронный ресурс] URL: http://www.ruscorpora.ru/corpora-intro.html
16. *Шахматов А.А.* Очерк современного русского литературного языка. М., 1941.
17. *Cabrera-Diego L. A., Torres-Moreno J.-M., Durette B.* Evaluating Multiple Summaries without Human Models: a First Experiment with a Trivergent Model. Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016.
18. Mystem+ [Электронный ресурс] URL: http://web-corpora.net/wsgi/mystemplus.wsgi/mystemplus/
19. *Levshina N.* How to do Linguistics with R. John Benjamins, 2015.
20. *Mehri A., Darooneh A. H.* The role of entropy in word ranking // Physica A: Statistical Mechanics and its Applications 390(s 18–19):3157–3163, 2011.
21. *Oakes M. P.* Statistical Measures for Corpus Profiling // Proceedings of the Open University Workshop on Corpus Profiling, London, UK, 2008.

22. *Rayson P., Garside R.* Comparing corpora using frequency profiling // Proceedings of the Comparing Corpora Workshop at ACL 2000. Hong Kong, 2000. P. 1–6.
23. *Zipf G.K.* The Psycho–Biology of Language: An Introduction to Dynamic Philology. Boston,1935.

References

1. Cabrera-Diego L. A., Torres-Moreno J.-M., Durette B. Evaluating Multiple Summaries without Human Models: a First Experiment with a Trivergent Model. *Natural Language Processing and Information Systems*: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016.
2. Dobrushina E.R., Litvintseva K.V., Pol'skov K.O., Khangireev M.A. 2011. 'Ot «abbata» do «analoia»: fragment lingvo-entsiklopedicheskogo slovaria russkoi khristianskoi leksiki'. *Vestnik PSTGU. Seriia III. Filologija,* no 3 (25). pp.119–148.
3. Dobrushina E. R. 'Slovar' khristianskoi leksiki: sostav slovnika'.*Vestnik PSTGU*, *Seriia III: Filologija,* 2012, no 3 (29), pp. 105–113.
4. D'iachenko G., prot. *Polnyi tserkovnoslavianskii slovar' s vneseniem v nego vazhneishikh drevnerusskikh slov i vyrazhenii,* Moscow, 1993 (reprint).
5. Zamkova V.V. *Slavianizm kak stilisticheskaja kategorija v russkom literaturnom jazyke XVIII v.,* Leningrad, 1975.
6. Kiynova Zh.K. 'Slavianizmy kak sredstvo stilizatsii v perevodakh religioznoi literatury'. *Vestnik Moskovskogo universiteta. Ser. 22. Teorija perevoda*. 2014. no 1. pp. 62–69.
7. Levshina N. *How to do Linguistics with R*, 2015.
8. Ledeneva V.V. 'Idiostil' kak sistema otnoshenii'. *Vestnik Tambovskogo universiteta. Seriia: Gumanitarnye nauki.* no 5 Vol. 23. 2001. pp. 12¬¬–13.
9. Litvintseva K.V. '«Tserkovnoslavianizm» kak lingvisticheskii termin'. *Vestnik Orlovskogo gosudarstvennogo universiteta.* 2015. no 6 (47). pp. 264–267.
10. Litvintseva K.V. 'Osobennosti funktsionirovaniia trekh frazeologizmov s leksemoi Bozhii v religioznykh i svetskikh tekstakh'. *Vestnik PSTGU. Seriia III: Filologija.* 2014. no. 4 (39). pp. 67–81.
11. Lyashevskaya O.N., Sharov S.A. *Novyi chastotnyi slovar' sovremennogo russkogo jazyka (nonlain-versiia)*, available at: http://dict.ruslang.ru/freq.php? (01.09.2017)
12. Lyashevskaya O.N., Sharov S.A. *Chastotnyi slovar' sovremennogo russkogo iazyka (na materialakh Natsional'nogo korpusa russkogo iazyka)*, Moscow, 2009.

13. Mehri A., Darooneh A. H. The role of entropy in word ranking. *Physica A: Statistical Mechanics and its Applications* 390(s 18–19), 2011. P. 3157–3163.
14. Mystem+ available at: http://web-corpora.net/wsgi/mystemplus.wsgi/mystemplus/ (01.09.2017)
15. Oakes M. P. Statistical Measures for Corpus Profiling. *Proc. of the Open University Workshop on Corpus Profiling*, London, UK, 2008.
16. Polyakov A.E. *Grammaticheskii slovar' tserkovnoslavianskogo iazyka (po materialam korpusa)*, available at: http://feb-web.ru/febupd/slavonic/dicgram/ (01.09.2017)
17. Rayson P., Garside R. 'Comparing corpora using frequency profiling'. *Proceedings of the Comparing Corpora Workshop at ACL 2000.* 2000. P. 1–6.
18. Sedakova O.A. *Slovar' trudnykh slov iz bogosluzheniia: Tserkovnoslaviano-russkie paronimy*, Moscow, 2008
19. Semenov P.A. 'Problema klassifikatsii stilisticheskikh funktsii slavianizmov (diakhronicheskii aspekt)'. *Vestnik Novgorodskogo gosudarstvennogo universiteta. Seriia: Gumanitarnye nauki.* 1998. no 4. pp. 134–138.
20. 'Chto takoe Korpus?' *Natsional'nyi korpus russkogo iazyka,* available at: http://www.ruscorpora.ru/corpora-intro.html (01.09.2017)
21. Shakhmatov A.A. *Ocherk sovremennogo russkogo literaturnogo iazyka.* Moscow, 1941.
22. Ulukhanov I.S. 'Tserkovnoslavianskii iazyk russkoi redaktsii: sfera rasprostraneniia i prichiny evoliutsii'. *Issledovaniia po slavianskim iazykam,* 2003, no 8, pp. 1–26.
23. Zipf G.K. *The Psycho–Biology of Language: An Introduction to Dynamic Philology.* Boston,1935.