Некоторые статистические характеристики корпуса церковнославянских богослужебных текстов

иеромонах Пантелеимон (Королев П. С.)
pantlmn@gmail.com

26 ноября 2018 г.

Аннотация

В статье рассматриваются...

Содержание

1. Предыстория

В 2013 году на конференции «Современная православная гимнография» в Институте русского языка им. В. В. Виноградова РАН мною был прочитан доклад «Гапаксы и иные статистические характеристики корпуса богослужебных текстов», однако, к большому сожалению, электронная версия доклада и все данные к нему были мною

утеряны и доклад так и не был опубликован. Исследование пришлось провести заново.

2. Содержание и размер корпуса

Для исследования были взяты тексты книг, содержащих богослужебные тексты:

1. богослужебное Евангелие,
2. богослужебный Апостол,
3. Псалтирь следованная,
4. Октоих,
5. Ирмологий,
6. Минея общая,
7. Минея месячная,
8. Триодь Постная,
9. Триодь Цветная,
10. Требник,
11. Часослов,
12. Служебник,
13. Молитвослов.

Электронные версии текстов книг были взяты с сайта orthlib.ru, созданном трудами священника Владимира Шина и М. Ю. Шин. Принципы выбора конкретных изданий для оцифровки нигде явно не прописаны, но преимущественно это московские издания конца XIX — начала XX века.

С сайта orthlib.ru также были взяты следующие тексты для включения в дополнительный корпус текстов:

1. Библия,
2. Типикон,
3. Акафистник,
4. «Алфавит духовный»,
5. «Добротолюбие»,
6. Минея праздничная,
7. Пролог,
8. Правила святых апостол,
9. Канонник,
10. разные последования.

Объем основного корпуса составил 2.6 млн словоупотреблений, вместе с дополнительным корпусом — 4.7 млн словоупотреблений.

Количество различных словоформ в основном корпусе — 95 тыс., вместе с дополнительным корпусом — 173 тыс.

А. Е. Поляков на материале корпуса подготовил грамматический словарь церковнославянского языка[1]. Словарь содержит 150 тыс. словоформ, группирующихся в 35 тыс. лексем. При этом лишь 22 тыс. из этих лексем встречаются в основном корпусе.

Можно выстроить как словоформы (рис. 1), так и лексемы (рис. 2) по частотности и увидеть, что соблюдается закон Ципфа[2]: количество употреблений q обратно пропорционально рангу r слова. Для наглядности сделаем обе шкалы на графике логарифмическими.

3. Наиболее часто употребляемые слова

Посмотрим на наиболее частотные лексемы в корпусе.

Вот двадцатка наиболее частых существительных (включая имена собственные): бо́гъ, госпо́дь, сла́ва, хрїсто́съ, гла́съ, о҃те́цъ, дꙋ́ша, пѣ́снь, богоро́диченъ, і҃рмо́съ, дѣ́ва, сло́во, дꙋ́хъ, вѣ́къ, богоро́дица, земла́, сы́нъ, моли́тва, свѣ́тъ, вѣ́ра.

Вот двадцатка наиболее частых прилагательных: свѧты́й, боже́ственный, бо́жїй, преподо́бный, чи́стый, благослове́нный, подо́бный, блаже́нный, честны́й, вѣ́рный, хрїсто́въ, небе́сный, пречи́стый, вели́кїй, госпо́днїй, мно́гїй, благі́й, ве́лїй, пра́ведный, сла́вный.

Вот двадцатка наиболее частых глаголов: бы́ти, пѣ́ти, ра́доватиса, глаго́лати, вопїа́ти, прїа́ти, моли́ти, спасти҃, ви́дѣти, роди́ти, ꙗви́тиса, прїити҃, ꙗ҃мѣ́ти, моли́тиса, рещи҃, благослови́ти, ꙗ҃зба́вити, дарова́ти, превозноси́ти, сотвори́ти.

Ничего удивительного в этих наборах нет, хотя, конечно, обращает на себя внимание наличие литургической терминологии (гла́съ, пѣ́снь, богоро́диченъ, і҃рмо́съ, ...)

4. Редкие словоформы и лексемы

Рассмотрим словоформы и лексемы, встречающиеся в основном корпусе лишь однажды — это так называемы гапаксы (hapax legomenon).

[1] http://dic.feb-web.ru/slavonic/dicgram/

[2] Alain Lelu. Jean-Baptiste Estoup and the origins of Zipf's law: a stenographer with a scientific mind (1868–1950) // Boletín de Estadística e Investigación Operativa. — 2014. — Т. 30, № 1. — С. 66–77.

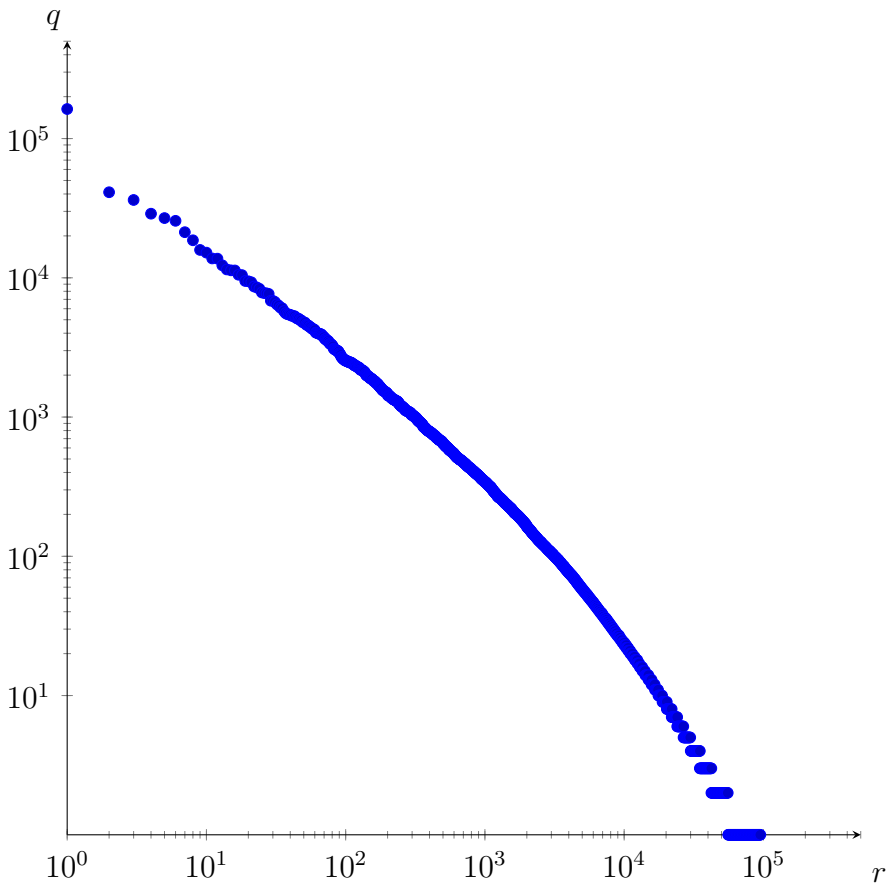

Рис. 1: Распределение словоформ по частотности.

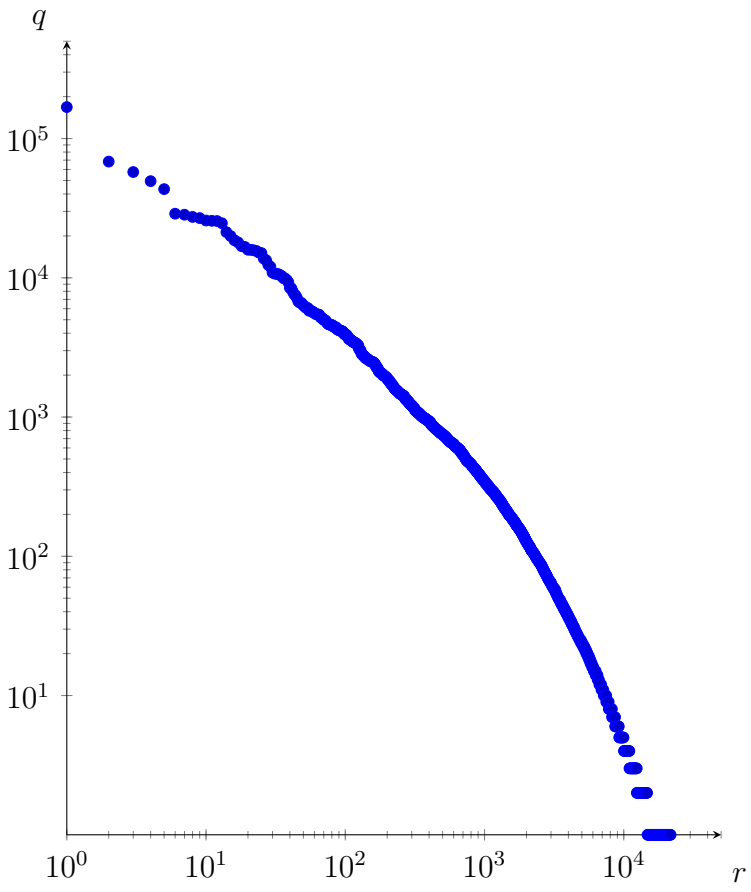

Рис. 2: Распределение лексем по частотности.

Редких словоформ на 1000 слов		Редких лексем на 1000 слов	
Богослужебный Апостол	54.00	Требник	9.44
Триодь Постная	52.05	Триодь Постная	7.50
Требник	43.90	Богослужебный Апостол	7.23
Служебник	41.13	Служебник	7.21
Богослужебное Евангелие	40.80	Богослужебное Евангелие	5.74
Следованная Псалтирь	33.18	Триодь Цветная	4.36
Минея месячная	31.96	Минея месячная	4.15
Триодь Цветная	31.01	Следованная Псалтирь	3.93
Молитвослов	30.29	Молитвослов	2.90
Октоих	27.40	Октоих	2.38
Минея общая	22.53	Минея общая	2.00
Ирмологий	18.59	Ирмологий	1.93
Часослов	12.86	Часослов	0.92

Таблица 1: Богослужебные книги и их лексическая сложность.

Среди словоформ таковых найдётся 39 тыс. (или 41% словоформ), а среди лексем — 7 тыс. (31% лексем).

Как подсказывает интуиция, сложность восприятия текста тесно связана с количеством редких слов, встречающихся в нём. Попробуем отранжировать богослужебные книги по сложности восприятия. Для этого возьмём не гапаксы, а словоформы и лексемы, встречающиеся в корпусе не более 3 раз. Назовём их «редкими».

Результаты см. в таблице 2. Различия при сортировке по разным параметрам незначительны.

1.	Требник. Алфавит имен	84.92
2.	Апостол. Сказание св. Епифания о 12 апостолах	41.22
3.	Апостол. Сказание Дорофея об избрании 70 апостолов	34.48
4.	Требник. Последование молебна больному перед хирургическим действием	29.20
5.	Триодь Постная. Сырная седмица. Суббота	29.01
6.	Требник. Молитва умирительная во вражде сущих	28.17
7.	Апостол. Краткое содержание апостольских Деяний по главам	26.32
8.	Служебник. Известие учительное	24.21
9.	Минея Общая. О знамениях владычних, и богородичных праздников, и святых	24.15
10.	Минея. Июнь. Неделя 2 по Пятидесятнице	23.22
11.	Требник. Последование о исповедании	22.75
12.	Триодь Цветная. Светлая Седмица. Пятница	22.57
13.	Требник. Чин освящения колесницы	19.80
14.	Требник. Молитва на всякую немощь	19.61
15.	Требник. Чин освящения храма	19.14
16.	Требник. Чин на нивах от вредителей	18.54
17.	Триодь Постная. Неделя о мытаре и фарисее	18.35
18.	Требник. Молитва запрещательная св. Василия над страждущими от демонов	17.54
19.	Триодь Цветная. Пятидесятница	17.09
20.	Триодь Постная. Седмица вторая. Пятница	15.75

Таблица 2: Богослужебные последования и их лексическая сложность (редких лексем на 1000 слов).