

Поляков А.Е.  
НПБ им. К.Д. Ушинского, Москва

## Корпус церковнославянских текстов: проблемы орфографии и грамматики<sup>1</sup>

### 1. Корпус.

В 2012 году впервые в истории мировой науки был создан и открыт для широкого доступа корпус церковнославянских текстов (<http://ruscorpora.ru/search-orthlib.html>). Этот корпус создан в рамках проекта «Национальный корпус русского языка» и открывает собой раздел исторических корпусов русского языка, которые должны охватить историю его развития от XI до XVIII века. Церковнославянский язык по своим лингвистическим свойствам ближе всего к богослужебному языку XVII–XVIII века, когда в основном была оформлена каноническая форма современных богослужебных текстов.

Церковнославянский язык, несмотря на свою старославянскую языковую основу, занимает важнейшее и естественное место в истории русского языка. Он всегда функционировал параллельно с русским и оказал заметное влияние на формирование русского литературного языка, в котором не только многие слова, но даже некоторые грамматические формы имеют церковнославянское происхождение, например, действительные причастия (*делающий, делавший*) и формы превосходной степени (*сильнейший*).

Церковнославянский язык, являющийся важнейшей частью русской культуры, до сих пор не имеет адекватного научного описания, отвечающего современному уровню. Существующие грамматики и словари носят в основном исторический или нормативный характер и описывают некоторую идеальную картину, которая часто не соответствует фактическому состоянию языка, отраженному в текстах. Без обращения к реальным текстам невозможно выяснить, как функционирует конкретное слово, как оно изменяется по формам, какие варианты написания имеет и т.д.

Адекватное научное описание церковнославянского языка должно носить не прескриптивный (нормативный), а дескриптивный (описательный) характер, опирающийся на реальное употребление. Единственной надежной основой для создания такого описания является **корпус** текстов, снабженных специальной разметкой (грамматической, структурной, метатекстовой). Корпус отличается от простого собрания текстов (библиотеки) именно наличием разметки, которая делает возможным поиск слов и словосочетаний по различным критериям (лемма, грамматические признаки, жанр текста и др.) и решение других задач. Корпус является важнейшим инструментом для исследований, который позволяет получить реальную информацию о лексике, грамматике и словоупотреблении из текстов.

---

<sup>1</sup> Данное исследование выполнено при поддержке РГНФ (проект 12-04-12045 «Электронная справочно-информационная система «Грамматический частотный словарь церковнославянского языка») и Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» (проект «Развитие корпуса церковнославянских текстов»).

Автор данной статьи является основным разработчиком церковнославянского корпуса. Кроме того, активное участие в создании корпуса принимали Е.Р.Добрушина (ПСТГУ), А.Г.Кравецкий (ИРЯ РАН), А.И. Зобнин (Яндекс), и ряд сотрудников ПСТГУ, ИРЯ РАН и компании «Яндекс».

## 2. Тексты и метатекстовая разметка.

Церковнославянский корпус включает около 1250 текстов, которые охватывают все основные типы и жанры церковнославянской литературы (богослужebные, святоотеческие, писание, типикон, церковное право). Корпус имеет объем около 4.6 миллиона словоупотреблений и включает около 150 тыс. различных словоформ, которые группируются примерно в 30 тыс. лексем. Корпус такого объема вполне репрезентативен с точки зрения охвата лексики и различных жанрово-тематических групп текстов. Все тексты подготовлены в едином формате и снабжены метатекстовой и грамматической разметкой.

Основным источником текстов для корпуса явилась «Библиотека святоотеческой литературы» (<http://orthlib.ru>). На данном ресурсе собраны результаты титанической работы по оцифровке и переводу в текстовый формат основных церковнославянских книг, за что создатели корпуса приносят искреннюю благодарность его создателям.

К сожалению, тексты в том виде, как они представлены в библиотеке, не вполне пригодны для корпуса. Во-первых, они представлены в нестандартной кодировке HP, которая удобна для набора на клавиатуре, но неудобна для чтения и обработки. Во-вторых, тексты не имеют лемматизации и грамматической разметки, что делает невозможным лексический поиск. Поэтому все тексты, взятые из библиотеки, были существенно переработаны и дополнены для нужд корпуса.

Метатекстовая разметка – это параметры, которые характеризуют текст в целом, а не отдельное слово. Сюда относятся следующие параметры текста:

- 1) Заголовок текста (иногда включает название книги).
- 2) Количество слов.
- 3) Жанр текста.
- 4) Период создания текста.
- 5) Перевод или оригинал.

Жанрово-тематическая классификация текстов ориентирована, прежде всего, на массового пользователя, а не на узкого специалиста, и строится исходя из принципов практического удобства и понятности. Выделяются следующие жанры текстов (с указанием % слов в корпусе):

- а) библия (17%): ветхий завет, новый завет, служба (богослужebное Евангелие);
- б) служба (64%): акафист, ирмологий, минея, октоих, служебник, третник, триодь, часослов;
- в) типикон (5%);
- г) святоотеческий (10%);
- д) право («Книга Правил святых апостолов» = 1%);
- е) научный («Ифика иерополитика» = 0.5%).

Период создания текста определяется условно, поскольку для многих книг невозможно точно определить дату создания, историю редактирования и дату последней правки. Дата издания позволяет определить только верхнюю границу, поскольку некоторые книги могли неоднократно редактироваться на протяжении XVII–XIX веков. В итоге выделяются следующие периоды:

- а) стандарт (XIX–XX = 86%);
- б) архаика (XVII–XVIII – «Пролог», «Добротолубие» = 2%);
- в) 20 век (3%);
- г) гибридный (8%, например, «Алфавит Духовный»).

Таким образом, основу корпуса составляют современные богослужebные тексты (60%), однако представлены и остальные жанры и периоды. Отдельно стоит старообрядческий «Пролог», который резко отличается от остального массива архаичным языком и орфографией.

### 3. Орфография и кодировка.

Каноническая церковнославянская орфография, ориентированная на точное типографское представление текста, включает множество избыточных символов, имеющих одинаковое фонетическое значение:  $u=i=v$ ,  $\phi=\theta$ ,  $o=o=w$ ,  $e=\epsilon$ ,  $ia=\lambda$ ,  $ou=uk$ ,  $\bar{w}=om$ ,  $s=z$ ,  $\xi=ks$ ,  $\psi=nc$ ,  $\bar{v}='$  (паерок), ударение (острое=тупое=облеченное).

В церковнославянском письме (в отличие от русского) некоторые пары символов используются для различения лексем и грамматических форм:

- в греческих словах сохраняется этимологически правильное написание букв  $u-i-v$ ,  $o-w$ ,  $\phi-\theta$ ,  $\xi$ ,  $\psi$ , например: *августъ—авраамъ*, *акундинъ*, *иквна*, *твхвнъ*, *Ѡвміамъ*;

- буквы  $u-i$  обычно распределены позиционно ( $i$  – перед гласным), однако используются для различения нескольких корней: *миръ* (спокойствие) vs. *міръ* (вселенная), *віно* vs. *вина*;

- буквы  $w$ ,  $\bar{w}$  используются для префиксов  $o(\bar{o})$ ,  $om$  и противопоставляются обычному  $o$ : *вблачати* vs. *облакъ*, *общій*; *всмотрити* vs. *осмь*; *вѣчатися* vs. *отчій*, *отрокъ*; в некоторых словах имеется колебание  $w/o$ : *вобразъ/образъ* (но *вбразовати*);

- буквы  $o-w$  используются для различения форм: единственное vs. множественное число (*милости* vs. *милѡсти*, *домомъ* vs. *домѡмъ*), винительный vs. родительный падеж в адъективном склонении (*новаго* vs. *новагѡ*, *моего* vs. *моегѡ*), прилагательное vs. наречие (*сильно* vs. *сильнѡ*) и в некоторых других случаях;

- для различения омонимичных форм единственного vs. множественного/двойственного числа также используются буквы  $e$  vs.  $\epsilon$  и острое vs. облеченное ударение (камора): *конемъ* vs. *конѣмъ*, *іерее* vs. *іерее̇*, *елени* vs. *елени̇*, *раба́* vs. *раба̇*, *рабу́* vs. *рабу̇*, *рабы́* vs. *рабы̇*.

Некоторые пары символов имеют нулевую или минимальную различительную способность:

- придыхание ставится автоматически над начальной гласной слова;
- диграф  $ou$  пишется в начале слова, лигатура  $uk$  в середине и конце;
- $ia$  пишется в начале слова,  $\lambda$  в середине и конце, за исключением корней *іазык-* (народ) vs. *азык-* (орган);

- $\epsilon$  (широкое) пишется в начале слова,  $e$  (узкое) в середине и конце, за исключением различения некоторых флексий единственного vs. множественного/двойственного числа (см. выше);

- $o$  (широкое) пишется в начале слова и корня в сложных словах (*отець*, *праотець*),  $o$  (узкое) в прочих случаях.

В результате тщательного анализа церковнославянской графики для представления корпуса в интернете была выработана более простая орфографическая система, которая сохраняет наиболее существенные языковые различия, но не пытается имитировать точный типографский вид текста. В этой орфографии отсутствуют некоторые символы с малой или нулевой различительной способностью (придыхания, различие  $ou/uk$ ,  $ia/\lambda$ ), однако сохраняются символы, связанные с различением лексем или грамматических форм ( $o-w$ ,  $u-i-v$ ,  $\phi-\theta$ ,  $z-s$ ). Кроме того, были выработаны правила перевода текста из этой орфографической системы в современный вид для удобства поиска.

Словоформы в корпусе представлены в том виде, в котором они встречаются в тексте, а леммы даются в унифицированном написании, в котором не используются избыточные буквы, а титла раскрыты (например, словоформа *мл̣ть* относится к лемме *милость*).

Для поиска в корпусе можно использовать три орфографические системы:

1) Точная – служит для поиска словоформ, сохраняет максимально точный вид словоформы (*e=ε, o=ο, титла, ударения*) с учетом вышеуказанных отличий от канонической орфографии.

2) Упрощенная – служит для поиска лексем, сохраняет основные лексические оппозиции (*o=w, u=i=v, φ=θ, z=s*), но игнорирует различия, не служащие для различения лексем (*e=ε, o=ο, титла*).

3) Модернизированная – служит для поиска лемм и словоформ в современной орфографии, включает только современные буквы (*e=b, u=i=v, φ=θ, z=s, o=w*).

Точная и упрощенная орфография ориентирована на специалистов, которые знают все тонкости церковнославянской орфографии и хотят получить наиболее точный результат. Модернизированная орфография рассчитана на широкий круг людей, которые хотят искать в корпусе, но не знают, как точно пишется некоторое слово (*лѣто* или *лето, ікѡна* или *икона*). Заметим, что модернизированная орфография просто заменяет буквы, но не переводит слово в современное написание (*езеро ≠ озеро, дѣлати ≠ делать, аггль ≠ ангел*).

Независимо от того, в каком виде слово введено в запрос, результаты поиска всегда выдаются в точной орфографии.

Основные соответствия между орфографическими системами сведены в таблице.

Символ	Каноническая (типографская)	Точная (словоформы)	Упрощенная (леммы)	Модернизированная
придыхание	(есть)	(нет)	(нет)	(нет)
букво-титла	млѣть, блгѣть	млѣть, блгѣть	милость, благодать	млѣть, блгѣть
паерок	бѣзѣзычнѣй	бѣзѣзычнѣй	бѣзѣзычнѣй	
титла	мѣти, аггль, блгѣть	мѣти, аггль, блгѣть	мати, ангель, благодать	мти, аггль блгѣть
оу- / -ук-	оучити, научити	учити, научити	у	у
іа- / -а-	іати, вбѣати	яти, вбѣяти	я	я
ѣ	сѣль	ѣ	ѣ	е
ε / -ε-	εзеро, кон+εмь	ε (езеро)	ε (езеро)	ε
ο / -ο-	οтець, праοтець	ο (отець)	ο (отець)	ο
w во флекс.	раб+wmь	раб+wmь	(нет)	ο
w в корне	ідwль, іkwна	w	w	ο
w в префиксе	wбразь/образь	w	w	ο
ѡ	ѡдати, ѡати	ѡдати, ѡяти	wtдати, wtѣяти	от-
і+согл.	ідwль, мѣрь, вѣна	і	і	и
і+глас.	знаніе	і	і	и
и	мирѣ	и	и	и
v ѡ	мѡро, еѡсѡгнѣй	v	v	и / в
θ	анѡраѣь	θ	θ	ф
ѣ	анѡраѣь	ѣ	ѣ	кс
ψ	апокалѣψісѣь	ψ	ψ	пс
s	сло, smіа	s	s	з

Церковнославянские тексты в корпусе представлены в стандартной кодировке Unicode. Нестандартные кодировки, которые применяются при наборе текстов в типографиях (HP, UCS), очевидно, не могут использоваться в интернете. Нам пришлось решать нетривиальную задачу по перекодировке из HP в Unicode, чтобы сохранить максимум значимых различий, присутствующих в исходной кодировке, однако пришлось пожертвовать некоторыми менее значимыми деталями. Дело в том, что некоторые важные символы (*ia, uk, букво-титла*) появились в стандарте Unicode

относительно недавно (в версии 5.2) и отсутствуют в большинстве доступных шрифтов. При просмотре в интернете браузер (по крайней мере, Mozilla Firefox) автоматически подгружает нужный шрифт и текст выглядит очень красиво, похоже на печатное представление. Однако, если выделить текст и скопировать его в редактор, некоторые символы будут отображаться как пустые квадраты, если на компьютере установлены старые версии стандартных шрифтов. Чтобы не отталкивать пользователей, которые не хотят устанавливать специальные шрифты, мы пожертвовали некоторыми символами: *ou/uk* заменили на обычное *y*, *ia/а* заменили на обычное *я*. Мы считаем, что лучше немного упростить текст, но оставить его читаемым, поскольку буквы *y* и *я* весьма частотны и их потеря существенно затруднит понимание текста (в отличие от потери редких букв – *Ѡ, ѡ, Ѣ, ѣ, Ѥ, ѥ*). В будущем, когда стандартные шрифты будут включать полный набор славянских символов, можно будет восстановить замененные буквы и тем самым приблизить вид текста в корпусе к канонической орфографии.

#### **4. Грамматическая разметка.**

Корпус отличается от простого собрания текстов наличием лингвистической разметки (грамматической, синтаксической, семантической и т.д.) и возможностью поиска по этой разметке. Для языков с богатой морфологией прежде всего необходима грамматическая разметка, которая обеспечивает возможность поиска слов по лемме (словарной форме) и грамматическим признакам. Поиск по точной словоформе (даже с усечением или шаблонами) абсолютно недостаточен для церковнославянского языка, в котором некоторые классы лексем имеют по несколько десятков словоформ (например, глаголы, включая формы причастий).

Грамматическая разметка для корпуса включает следующие задачи:

- 1) лемматизация – приведение словоформы к лемме (словарной форме) и определение ее грамматических признаков (часть речи, род, вид, переходность);
- 2) грамматический анализ – определение грамматических признаков словоформы (падеж, число, время, лицо, наклонение).

Эти задачи могут решаться одновременно, но с разной степенью полноты и точности. Пользователям прежде всего нужен поиск по лемме и части речи, в меньшей степени – по другим грамматическим признакам. Кроме того, в церковнославянском очень распространена грамматическая омонимия между словоформами одной лексемы, поэтому точно определить их грамматические признаки часто невозможно.

Существующие грамматики и словари церковнославянского языка не дают полной картины словоизменения. Граматики приводят парадигмы для наиболее частотных слов, а также отдельные примеры вариативных форм. Словари обычно включают наиболее частотные или важные слова, а грамматическая информация дается фрагментарно или не дается вообще. В результате нередко невозможно выяснить, какие реальные формы имеет данное слово, а спорные вопросы остаются без ответа (примеры см. в п. 5). В таких случаях единственным достоверным источником является корпус.

Грамматическая модель словоизменения церковнославянского языка не задается априорно на основе существующих грамматик и словарей, а выводится эмпирически на основе анализа данных корпуса. Модель словоизменения включает в себя два основных компонента:

- 1) грамматический словарь,
- 2) таблица словоизменительных типов (парадигм).

Грамматический словарь представляет собой список лексем с приписанной им информацией о словоизменении. Каждая лексема в словаре содержит, как минимум, следующие параметры:

- лемма (словарная форма) и ее варианты (если есть);
- постоянные признаки лексемы (часть речи, одушевленность, переходность);

- код парадигмы и особенности словоизменения (нерегулярные словоформы);
- краткое толкование для устаревших и малопонятных слов (по необходимости).

Словоизменительный тип (парадигма) представляет собой список грамматических значений и соответствующих им грамматических форм, общий для некоторого множества лексем.

Парадигмы также не задаются априорно на основе существующих описаний, а выводятся из корпуса на основе анализа множества словоформ, имеющих однотипное соотношение между грамматическими формами. Таким образом, номенклатура парадигм получается значительно более детальной и может существенно отличаться от традиционных грамматик. Например, традиционное первое склонение (*рабъ*) распадается на 14 подтипов в зависимости от конечного согласного (парный твердый–мягкий, велярный, шипящий, йот), наличия беглого гласного и других особенностей.

Вот небольшой фрагмент таблицы парадигм в формальной записи.

Парадигма	N1t	N1t*	N1j	N1k	N1x	N1k*
Пример	рабъ	осель, сонъ	конъ, царь	отрокъ	духъ	свитокъ
Основа	раб+ъ	осе*л+ъ, со*н+ъ	кон+ъ, цар+ъ	отро(к ц ч)+ ъ	ду(х с ш)+ъ	свит(к ок ц ч)+ ъ
sg,nom	ъ	2ъ	ь	ъ	ъ	2ъ
sg,acc	=nom/gen	=nom/gen	=nom/gen	=nom/gen	=nom/gen	=nom/gen
sg,gen	а	а	я	а	а	а
sg,dat	у	у	ю	у	у	у
sg,loc	ѣ	ѣ	и	2ѣ	2ѣ	3ѣ
sg,ins	омъ	омъ	емъ	омъ	омъ	омъ
sg,voc	е	е	ю	3е	3е	4е
pl,nom/voc	и	и	и/ie	2ы	2и	3ы
pl,acc	ы/=gen	ы/=gen	и/=gen	и/=gen	и/=gen	и/=gen
pl,gen	wвъ/ъ^	wвъ/ъ^	ей	wвъ	wвъ	wвъ
pl,dat	wмъ	wмъ	емъ	wмъ	wмъ	wмъ
pl,loc	ѣхъ	ѣхъ	ехъ	2ѣхъ	2ѣхъ	3ѣхъ
pl,ins	ы	ы	и/ьми	и	и	и
du,nom/acc	а^	а^	я^	а^	а^	а^
du,gen/loc	у^	у^	ю^	у^	у^	у^
du,dat/ins	ома	ома	ема	ома	ома	ома

В этой таблице в первой строке указаны коды парадигм, в первом столбце – грамматические формы в виде набора признаков. Во второй строке даются примеры лемм в обычной записи, в третьей – в формате грамматического словаря. Чередования записываются как формулы вида (x|y|z), например, *отро(к|ц|ч)* означает, что основа имеет три варианта – *отрок*, *отроц*, *отроч*. Номер варианта используется при записи флексий, причем первый вариант считается основным. Например, флексия "у" требует первого варианта основы (*отрок+у*), флексия "2ъ" имеет форму "ѣ" и требует второго варианта (*отроц+ъ*), флексия "3е" требует третьего варианта (*отроч+е*).

Грамматический словарь и модель словоизменения создаются эмпирически и итеративно. Сначала из корпуса генерируется список словоформ и делается его первичная проверка и чистка (исправляются явные ошибки). Затем для частотных слов делается ручная лемматизация, определяются типичные шаблоны и строится первичная модель словоизменения. Далее на основе шаблонов разобранных словоформ анализируются другие словоформы, затем автоматические разборы проверяются и правятся вручную, снова уточняется грамматическая модель, и так далее. На последнем

этапе разбираются варианты и уникальные словоформы и принимается решение о том, что с ними делать (исправить, включить в словарь, считать исключением, игнорировать).

В настоящее время создана пилотная версия грамматического словаря, которая используется для поиска в корпусе. Было проанализировано около 150 тыс. словоформ, которые в итоге были сведены примерно в 30 тыс. лемм. В результате ручной, а затем автоматической лемматизации удалось приписать грамматические разборы большинству словоформ. Часть словоформ была квалифицирована как ошибки (опечатки), а для некоторых редких форм лемму определить не удалось или она была определена гипотетически.

### 5. Вариативность.

В церковнославянских текстах имеется множество орфографических и словоизменительных вариантов, обусловленных тем, что тексты создавались в разное время и в разной языковой среде. Вот некоторые примеры.

1) В грамматиках дается список слов и корней, которые обычно пишутся под титлом: *аггль*, *апль*, *бгъ*, *сдь*, *млть*, *мти*, *стбий*, *члвкъ*. Возникает вопрос: всегда ли эти корни пишутся именно так (под титлом), а если нет, то как еще они могут писаться? Ответ на этот вопрос может дать только корпус. В результате анализа разных написаний в корпусе оказывается, что например, корень *апостол-* может писаться тремя способами: *апостол-* (полностью раскрыто), *апл-* (полностью сокращенно), *аптол-* (частично сокращенно); основа прилагательного *ангельскій* пишется тремя способами: *аггельск-*, *аггльск-*, *аггльск-*. Без корпуса выявление всех реально встречающихся вариантов было бы просто невозможно.

2) В грамматиках написано, что для различения омонимичных форм единственного и множественного/двойственного числа может использоваться замена букв ( $o \rightarrow w$ ,  $e \rightarrow \epsilon$ ) или замена острого ударения на облученное (камору). Таким образом, возникает конкуренция орфографических правил и непонятно, какое из них нужно применять в конкретном случае. Например, как будет множ. число от слов *милость*, *высота*: *милѡсти* ( $o \rightarrow w$ ) или *мѡсти* (камора), *высѡты* ( $o \rightarrow w$ ) или *высоты* (камора)? Ответ на этот вопрос может дать только корпус, который дает следующую статистику: *высѡты* (17 раз), *высоты* (9 раз), *высѡты* (1 раз), *милѡсти* (93 раз), *мѡсти* (2 раза). Из анализа других аналогичных примеров мы можем сделать вывод, что при конкуренции правил замена  $o \rightarrow w$  более предпочтительна, чем замена ударения.

3) В грамматиках написано, что слова с суффиксом *-тель* типа *дглатель* могут иметь особые формы именит. множ. на *-е* (*дглателе*, как *агаряне*) и на *-іе* (*дглателіе*, как *царіе*). Кроме того, они могут иметь стандартные формы на *-и* (*дглатели*). Естественно, возникает вопрос о том, какие из этих форм реально встречаются и с какой частотой. Решить подобный вопрос возможно только при помощи корпуса, который дает следующую статистику: *дглатели* – 36 раз, *дглателіе* – 18 раз, *дглателе* – 13 раз. Таким образом, для данного слова чаще всего употребляется стандартная форма, за ней идет форма на *-іе*, а затем на *-е*. Для получения достоверных выводов нужно проанализировать остальные слова с суффиксом *-тель*, а также другие слова с аналогичными особенностями (с суффиксом *-арь* и т.д.).

4) Церковнославянский язык является смешанным по своему грамматическому строю. Многие грамматические формы совпадают с русскими, а несовпадающие, происходящие от старославянских, нередко вытесняются эквивалентными русскими формами. Например, наряду со старославянской формой *на земли* встречается русская форма *на земль*; предложный падеж множественного числа от слов первого склонения имеет вариативные формы *сельхъ/селахъ*, *жилицяхъ/жилицихъ*, но только *моряхъ*, *мужахъ* (при отсутствующих *\*морихъ*, *\*мужихъ*). Вопрос о конкуренции

старославянских и русских форм изучен совершенно недостаточно, а для получения достоверных результатов нужен корпус.

**Литература.**

*Добрушина Е. Р., Поляков А. Е.* Корпус церковнославянского языка: возможности, методы создания, перспективы // Вестник ПСТГУ. Серия III. Филология. № 1 (31), 2013. – С. 32–44

*Алипий (Гаманович).* Грамматика церковнославянского языка. – Jordanville (N. Y.), 1964.

*Плетнева А. А., Кравецкий А. Г.* Церковнославянский язык. – М., 2001.

*Алексеев П.* Церковный словарь. Часть I–IV. – М., 1815–1816.

*Дьяченко Г.* Полный церковно-славянский словарь. – М., 1900.

Словарь церковно-славянского и русского языка, составленный Вторым отделением Императорской Академии наук. Том I–IV. – СПб., 1847.