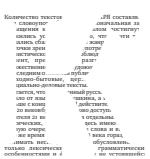
С. А. Оскольская

Корпус письменных текстов XIX века:

сферы употребления и жанровое разнообразие

Введение



оличество текстов 19 века в НКРЯ составляет 26 млн словоупотреблений. Первоначальная задача насыщения корпуса материалом достигнута, и сложились условия для того, чтобы эти тексты оказались сбалансированы в жанровом отношении и с точки зрения сферы их употребления. Основное стилистическое деление, соблюдаемое в настоящий момент, предполагает разграничение между художественными и нехудожественными текстами. К последним относятся публицистические, научные, обиходно-бытовые, церковно-богословские и официально-деловые тексты.

Считается, что современный русский язык ведет свое начало от языка А.С. Пушкина, а возможно, и еще раньше—с конца 18 века. И действительно, два текста—19 и 20 веков—будут одинаково доступны пониманию читателя 21 века, не считая отдельных, в основ-

ном лексических, элементов (здесь имеются в виду в первую очередь устаревшие слова и выражения). В то же время текст начала 18 века гораздо труднее понимать неспециалисту, и обусловлено это не только лексическими, но и грамматическими особенностями и фактом еще не устоявшейся к тому времени нормы.

Наличие в Национальном корпусе русского языка массива текстов 19 века позволяет прослеживать на протяжении двухсот лет развитие в русском языке того или иного явления, например, изменение в управлении какой-либо глагольной лексемы, развитие новых лексических значений, изменение грамматических характеристик слова (ср. колебания в роде у слов типа *рояль* и *лебедь*, склоняемость имен типа *кофий-кофе*, этапы освоения заимствований). Для таких наблюдений можно ранжировать тексты в поиске по приблизительному времени их создания. При этом правильное статистическое распределение ранних письменных текстов по жанрам и сферам употребления должно стремиться к тому, которое существовало в момент их создания.

2. Распределение текстов

На данный момент в Национальном корпусе русского языка насчитывается около 26 млн словоупотреблений в 1500 единицах текстов 19 века. Существенно, что единицы текстов могут значительно различаться по объему (ср. роман «Война и мир» Л.Н. Толстого и образцы деловой переписки, состоящие порой из нескольких строк).

Преобладающая часть художественных текстов была собрана еще в период с 2003 по 2005 гг. (около 20 млн словоупотреблений). Последние три года корпус 19 века пополнялся в основном учебно-научной литературой и текстами публицистического и обиходно-бытового характера[1]. С 2006 по 2008 гг. собрано более 6 млн. словоупотреблений. Распределение текстов по сферам функционирования и по жанрам представлено в таблицах 1 и 2.

---

[1] Сбор данных финансировался из проекта «Сбор и обработка данных в формате Национального корпуса русского языка», поддержанного программой Президиума РАН «Русский язык, литература и фольклор в информационном обществе: формирование электронных научных фондов» ИМЛИ ЗОИФ (руководитель проекта – М.Д. Воейкова, ИЛИ РАН).

Таблица 1.

Сфера функционирования	% словоупотреблений
художественная	56,3%
публицистика	24,4%
учебно-научная	12%
обиходно-бытовая	4,6%
церковно-богословская	2%
официально-деловая	0,7%

Таблица 2.

Жанр текста	% словоупотреблений
нежанровая проза	74%
историческая проза	8,7%
документальная проза	5,4%
драматургия	5,2%
юмор и сатира	2,7%
приключения	2,1%
фантастика	1%
детская	0,6%

Как видно из приведенных данных, существует необходимость в увеличении доли обиходно-бытовых и официально-деловых текстов. Понятно, однако, что и в момент создания процент таких текстов был существенно ниже, нежели процент художественных и публицистических произведений, составлявших основной круг чтения в 19 веке. Сравнение приведенных данных с данными 2005 г. (см. статью Н. Л. Дич в сборнике «Национальный корпус русского языка 2003–2005», с. 90) показывает, что соотношение текстов по сферам функционирования за последние три года выравнивалось в сторону сбалансированности: если в 2005 г. доля художественных текстов составляла 66%, то сейчас, три года спустя, она снизилась до 56,3%. Значительно (с 7,2% до 12%) повысилась доля учебно-научных текстов. Процентная же доля обиходно-бытовых и официально-деловых текстов повысилась незначительно (на 0,3 и 0,5% соответственно).

Основу нежанровой художественной прозы составляют романы (56% словоупотреблений), повести (19%), рассказы (12%) и очерки (10%).

Учебно-научная сфера функционирования включает в себя тексты различных научных областей. Распределение научных текстов по тематике представлено в табл. 3.

Таблица 3.

Тематика текста	**% словоупотреблений**
политология (политика и общественная жизнь)	32%
религиоведение	15%
естественные науки	17%
философия	13%
филология	10%
математика	5%
психология	3%
право	1%

Многие политические тексты совмещают в себе черты научной и публицистической функциональной сфер, поэтому они и составляют столь значительную долю от общего числа научных текстов.

Естественнонаучная область представлена монографиями, статьями и заметками по биологии (работы А. Я. Данилевского, Н. Е. Введенского, И. И. Мечникова и др.), географии и геологии (работы Д. Н. Анучина), медицине (работы Ф. Ф. Эрисмана), химии (работы Н. Д. Зелинского, А. М. Бутлерова) и физике (работы П. Н. Лебедева). Большая часть трудов по математике принадлежит перу П. Л. Чебышева и М. В. Остроградского.

Среди авторов исторических работ можно назвать Н. М. Карамзина, В. Н. Татищева. Философия представлена трудами Л. М. Лопатина, Вл. Соловьева. Психология—работами В. М. Бехтерева. Правоведение—работами А. Ф. Кони.

Публицистическая сфера функционирования представлена трудами Л. Н. Толстого, К. Н. Леонтьева, Н. И. Новикова и других авторов. Самыми распространенными типами публицистических текстов оказываются мемуары (64%), статьи (23%) и очерки (8%).

Обиходно-бытовую сферу функционирования составляют такие типы текстов, как переписка (например, переписка П. И. Чайковского с Н. Ф. фон Мекк), дневники и записные книжки (например, дневник Д. М. Волконского 1812–1814 гг.) и различные записки и очерки.

В церковно-богословскую сферу функционирования входят следующие типы текстов: беседа, житие, катехизис, молитва, поучение, проповедь и некоторые другие. Авторами большинства имеющихся в корпусе церковно-богословских текстов—не считая, конечно, Священного писания—являются архиепископ Иннокентий, Игнатий Брянчанинов, Л. Н. Толстой.

Официально-деловая сфера функционирования представлена различными приказами, докладами, манифестами, деловыми письмами и пр.

3. Источники текстов

Часть текстов была предоставлена в электронном виде издательствами, в частности, издательствами «Наука» и «Нестор-История». Некоторые отсканированные тексты взяты из проекта «Эго-документ в литературно-письменной традиции 19 века» (руководитель В. Н. Калиновская, ИЛИ РАН), который проводится в рамках Программы фундаментальных исследований Секции языка и литературы ОИФН РАН «Русский язык, литература и фольклор в информационном обществе: формирование электронных научных фондов».

Для большинства функциональных сфер необходимо отметить труднодоступность текстов 19 века. По сравнению с художественной литературой, крайне мало текстов научного, официально-делового или, например, обиходно-бытового характера переведено в электронный вид и выложено на сайтах в сети Интернет. Поэтому многие тексты приходится сканировать или фотографировать со старых изданий избранных трудов ученых 19 века и с книг, в которых опубликованы некоторые документы и другие архивные материалы. Так, например, благодаря сканированию книги «Бородино: Документальная хроника» (М.: «Российская политическая энциклопедия» (РОССПЭН), 2004) Национальный корпус пополнился документами, затрагивающими тему Бородинского сражения: приказами, докладами, отчетами, обзорами, деловыми письмами и пр.

Несмотря на крайне небольшое количество нехудожественных текстов, опубликованных в сети Интернет, все-таки можно отметить несколько сайтов, послуживших источниками отдельных текстов. Речь идет о специализирующихся исторических и литературных сайтах, на которых опубликованы различные архивные мате-

риалы 18–20 вв.: Фундаментальная электронная библиотека «Русская литература и фольклор» (*http://feb-web.ru/*), сайт «Русские мемуары» (*http://memoirs.ru*), сайт «Восточная литература—библиотека текстов Средневековья», на котором собраны также многие российские документы 18 и 19 веков (*http://www.vostlit.info/*) и некоторые другие.

Все тексты были вычитаны и проверены на наличие ошибок сканирования или набора и отформатированы по единым правилам.

4. Проблема старой орфографии

Некоторые тексты попадали к нам в дореволюционной орфографии. Поскольку многие тексты были введены в Национальный корпус уже в новой орфографии, было принято решение переводить все тексты в современную орфографию в соответствии с реформой 1918 г. Так, в конце слов убраны все знаки Ъ, буквы Ѣ, Ѳ, V, І заменены на Е, Ф, И, И соответственно. Старые окончания прилагательных, причастий и местоимений заменены на современные (*-аго* на *-ого*, *-ыя* на *-ые* и др.). Приставки, заканчивающиеся на *-з-*, в соответствующих фонетических условиях вместо *-з-* получали *-с-*. Местоимения *оне* и *ея* заменялись на *они* и *ее*. Также были выполнены и некоторые другие изменения согласно реформе 1918 года.

В текстах были оставлены те отклонения от правил, которые никак не отражены в реформах орфографии и являются скорее особенностью авторского стиля или времени, нежели проявлением нормы русского языка, если о таковой вообще можно говорить по отношению к 19 веку. Например, были оставлены такие формы, как *генваря*, *повидимому* или *чорт*. Безусловно, это затрудняет поиск (в некоторых случаях нахождение словоформы возможно только при поиске точных форм), однако позволяет сохранить особенности текста 19 века, которые могут быть важны при проведении различных лингвистических исследований. Предполагается, что в дальнейшем будет проведена модификация поисковой программы, что позволит учитывать при запросе различия в орфографии отдельных слов и производить их отбор как в старой, так и в новой орфографии, а при необходимости и совместный поиск.