Н. Л. Дич

О ТЕКСТАХ XIX ВЕКА В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА

Состав корпуса текстов XIX века

Корпус XIX века охватывает период от начала XIX до начала XX века: самые ранние тексты представлены сочинениями Н. М. Карамзина, В. Т. Нарежного и В. А. Жуковского, самые поздние — произведениями Чехова, Андреева, Бунина, Короленко и др.

Так же, как и корпус современных текстов, корпус XIX века включает в себя тексты различных функциональных сфер, жанров и типов, однако распределяются тексты по функциональным сферам, жанрам и типам в двух корпусах по-разному.

Различия в распределении текстов объясняются несколькими причинами, и в первую очередь тем, что многие типы текстов XIX века практически недоступны в электронном виде (или в таком виде, который можно было бы легко перевести в электронный). Так, например, достать образцы личных писем XIX века — задача намного более трудоемкая, чем найти образцы современной личной переписки. Дошедшие до нас письма XIX века принадлежат главным образом перу писателей, философов, политиков и ученых, и зачастую эти тексты по стилю и тематике граничат с художественной литературой и публицистикой, то есть не являются бытовой перепиской в чистом виде.

Труднодоступностью источников объясняется также и то, что на сегодняшний день в корпусе XIX века плохо представлены деловые и газетно-журнальные тексты (конкретные цифры буду приведены ниже).

Распределение текстов по типам и тематике обуславливается и спецификой эпохи: так, например, в корпусе XIX века отсутствуют тексты производственно-технической и рекламной сферы.

Распределение текстов по функциональным сферам, жанрам и тематике

В таблицах 1-3 приводится статистика корпуса XIX века на конец октября 2005 года. Общий объем корпуса — 22 млн. словоупотреблений.

Табл. 1 Распределение текстов по сферам функционирования.

сфера функционирования	**% словоупотр.**
художественная	66%
публицистика	18,9%
учебно-научная	7,2%
бытовая	4,3%
церковно-богословская	3,4%
официально-деловая	0,2%

Табл. 2 Распределение художественных текстов по жанрам.

жанр	**% словоупотр.**
нежанровая проза	79,2%
историческая проза	9,6%
автобиографическая проза	4,5%
юмор и сатира	2,9%
приключения	2,1%
фантастика	1%
детская	0,7%

Табл. 3 Распределение нехудожественных текстов по тематике.

тематика текста	**% словоупотр.**
политика и общественная жизнь	20,9%
религия	20,8%
наука	19,5%
философия	14,8%
частная жизнь	9,8%
армия и вооруженные конфликты	7,2%
искусство и культура	5,3%
другие темы	1,8%

РАСПРЕДЕЛЕНИЕ ПО ТИПАМ ТЕКСТОВ

Художественные тексты представлены главным образом романами, повестями, рассказами и очерками (56, 18, 11 и 10% словоупотреблений соответственно).

Основные типы публицистических текстов — статьи, мемуары и очерки (37, 36, 22% словоупотреблений соответственно).

Учебно-научная литература представлена монографиями и статьями по истории, философии, а также несколькими работами по естественным наукам, языкознанию и литературоведению.

Тексты бытовой сферы включают в себя переписку, дневники и записные книжки. Сюда же были отнесены черновики и наброски. Но, как уже упоминалось, значительная часть бытовых текстов не являются бытовыми в чистом виде. Как письма, так и многие дневники и записные книжки по стилю и тематике граничат с публицистикой (например, «Журнал 1813 года» А. И. Михайловского-Данилевского является одновременно дневником и мемуарами) или с художественной литературой (это случай отрывков и набросков художественных текстов).

Церковно-богословская сфера представлена в основном проповедями и статьями. Официально-деловая сфера — программами партий и постановлениями.

Источники текстов

При подготовке корпуса XIX века использовались сканированные тексты (в том числе из коллекции Машинного фонда ИРЯ им. В. В. Виноградова РАН) и профессионально подготовленные электронные версии, находящиеся в свободном доступе в сети Интернет — в частности, ресурсы РГБ (orel.rsl.ru), ФЭБ (feb-web.ru) и нек. др.

Все электронные версии, независимо от их происхождения, проверялись на наличие ошибок набора и сканирования.

Особенности метаразметки

Многие из проблем, которые приходится решать при подготовке текстов XIX века, связаны с метаразметкой. Одна из основных вопросов — это определение типа текста.

Поскольку многие жанры и типы текстов в XIX веке еще только формируются, применение современной классификации зачастую оказывается затруднительным.

В пограничных случаях, когда в критической литературе текст идентифицируется по-разному, используются двойные пометы. Так, например, «Капитанская дочка» Пушкина имеет в корпусе помету: роман|повесть. Имеются также тексты с пометами рассказ|повесть, очерк|рассказ, очерк|повесть и др.

Не вполне очевидно, в какой степени при определении типа текста XIX века можно опираться на его самоидентификацию, поскольку термин, использованный автором для идентификации текста, мог изменить свое значение. Например, современное понимание *романа* предполагает произведение довольно большого объема; в текстах XIX века это условие соблюдается далеко не всегда.

При описании текстов XIX века зачастую сложно провести границу не только между типами текстов, но и между функциональными сферами, в частности, между художественной литературой и публицистикой. Пограничное явление представляют собой, например, некоторые юмористические рассказы Тэффи, которые, несмотря на то, что они названы самим автором *рассказами,* могут быть отнесены к *фельетонам*, поскольку написаны на злободневные темы и не всегда имеют сюжет.

Также не всегда можно разграничить публицистику и научную литературу: тексты по политологии (напр., В. И. Ленин) или философии (напр., Вл. Соловьев) должны быть отнесены к научными в силу их тематики и используемой терминологии. При этом они характеризуются эмоциональной окраской, не вполне свойственной научному стилю.

Выше уже упоминались случаи текстов, стоящих на границе бытовой и художественной или бытовой и публицистической сфер.

В случаях, когда границу между функциональными сферами при разметке текста провести невозможно, также используются двойные пометы.

Проблемы старой орфографии и устаревших форм

Многие тексты, входящие в корпус XIX века, сохраняют особенности старой орфографии: встречаются такие написания, как *под устцы, времяна, поперег, по-тихоньку* и пр. В корпусе со снятой омонимией и с полной морфологической разметкой такие формы должны были бы, по-видимому, иметь помету «искаженная форма» (именно такая помета приписывается в корпусе современных текстов написаниям типа *фторник,* передающим особенности произношения). Однако пока полная морфологическая разметка не осуществлена, контексты, содержащие подобные написания, будут выдаваться пользователю только при поиске точных форм. Если же пользователь будет искать соответствующую лексему по ее

словарной форме, нестандартные написания останутся за рамками рассмотрения. Вопрос о том, следует ли сохранять подобные написания и обозначать их как «искаженная форма» при морфологической разметке, или же заменить их на написания, соответствующие современным правилам орфографии, пока не решен.

Наличие в текстах XIX века устаревших морфологических форм также представляет собой проблему для пользователя. Такие нестандартные с точки зрения современного языка формы, как *грустию, бревны, вышед* и пр. в текстах XIX века встречаются регулярно, однако — точно так же, как в случае с устаревшими написаниями — эти формы не выдаются пользователю при лексико-грамматическом поиске.

Например, если пользователь поставит перед собой задачу найти все контексты, содержащие формы слова «выйти», и задаст в окне лексико-грамматического поиска «выйти», он не получит ни одного из нескольких десятков имеющихся в корпусе контекстов с формой «вышед». Чтобы найти эти примеры, нужно задать «вышед» в окне поиска точных форм, но для этого пользователь должен ***знать***, что такая форма в языке существовала!

Один из возможных путей решения данной проблемы — подробная морфологическая разметка текстов, которая предусматривает использование пометы «аномальная форма». Учитывая значительное количество контекстов, содержащих устаревшие формы, подробная разметка хотя бы части текстов XIX века со снятием грамматической омонимии становится одной из первостепенных задач на последующих этапах работы над Национальным корпусом русского языка[1].

[1] *От редакции:* аналогичная проблема возникает при разметке текстов электронной коммуникации, размещаемых в Корпусе. Очевидно, этот вопрос для текстов обоего вида должен решаться единообразно.