А. А. Бурыкин, В. П. Захаров
A. A. Burykin, V. P. Zakharov

КОРПУС И ПРОБЛЕМЫ ГРАФИКИ И ОРФОГРАФИИ: НАБЛЮДЕНИЯ ИЗ ОПЫТА РАБОТЫ ПО СОЗДАНИЮ «БИБЛИОТЕКИ ЛЕКСИКОГРАФА»

CORPUS AND PROBLEMS OF GRAPHICS AND ORTHOGRAPHY: OBSERVATIONS FROM THE EXPERIENCE OF COMPILING OF THE «LEXICOGRAPHER LIBRARY»

Аннотация. Авторы статьи предлагают обсудить некоторые вопросы русской графики и орфографии в исторической перспективе, важные для создания корпусов текстов и работы с ними. Некоторые нерешенные проблемы русской орфографии создают серьезные проблемы при обработке текстов для корпуса и при их использовании в составе корпуса. Одновременно с этим значительные массивы электронных текстов позволяют извлекать и детерминировать в хронологическом отношении отдельные факты истории русской орфографии, не отраженные орфографическими словарями русского языка.

Abstract. The article suggests to discuss some questions of Russian writing system and spelling in historical perspective, important for creating corpora and for the practical work with them. Some unresolved problems of Russian spelling create serious problems at text processing. At the same time the considerable amount of electronic texts allows to define in chronological relation some facts of history of the Russian spelling, not reflected by spelling dictionaries of Russian.

1. «Библиотека лексикографа»

Настоящая статья является результатом работы над проектом «Библиотека лексикографа» – собранием русских текстов для лексикологических исследований и лексикографической практики, который создан и используется в

Словарном отделе Института лингвистических исследований РАН с 2008 г. В настоящее время «Библиотека лексикографа» включает более 42 тыс. текстов разных жанров и различной тематики. Суммарный объем ресурса – около 1,7 млрд. словоформ.

Работа над любым массивом текстового материала неизбежно соприкасается с вопросами графической репрезентации текста. Понятно, что решение этих вопросов во многом зависит от задач, на решение которых направлен корпус или массив текстов: так, ориентация корпуса на отражение современного состояния языка жестко требует выдерживания и единообразия графико-орфографических норм современного русского языка. Но и данная ситуация, казалось бы, несложная, отнюдь не проста – в современной русской графике, орфографии и пунктуации остается немало дискуссионных и нерешенных вопросов. Положение дел еще сложнее, если корпус, как «Библиотека лексикографа», должен охватывать тексты XVIII-начала XXI веков, при этом для разработчиков важно, чтобы в нем учитывалась аутентичная орфография текстов, позволяющая наблюдать процессы изменения в русской орфографии за последние три столетия не по словарям и нормативным сводам правил, а по текстам. Добавим к этому, что для ряда периодов истории русского литературного языка, охватывающих целые десятилетия, нормативные орфографические словари вообще отсутствуют, и как изучать характер норм орфографии, например, в интервале 1918-1955 гг. – не вполне понятно.

Отдельный вопрос – как получить электронные версии текстов, свободные от поновления орфографии, но он выходит за рамки обсуждения.

2. Проблема «буквы Ё»

Проблема «буквы Ё» становится камнем преткновения как для исследований на основе электронных версий текстов, так и для разработки шрифтов и программ для работы с текстами. Венцом всего здесь является «обезъяченный и объёшенный»

словарь В.И. Даля в современной версии. В то же время не только в нем, но и, например, в электронных версиях словаря Д.Н. Ушакова и С.И. Ожегова не только расставлена буква ё, отсутствующая в исходных печатных версиях этих словарей, но и введен новый порядок расположения слов – слова с Ё стоят после слов с Е, а не вперемежку с ними, как традиционно принято в русской лексикографии. Активное внедрение графического знака Ё (буквой в строгом смысле слова он не является из-за особенностей функционирования и самого характера действующих норм, провозглашающих факультативность) приводит к следующему: в громадном количестве «объёшенных» текстов знак Ё выбивается из кодировки и отражается в виде непонятных символов, которые не идентифицируются при поиске и препятствуют выявлению текстов со словами с расставленным Ё в тексте.

Поисковые инструменты большинства программ «разводят» знаки Е и Ё, в результате чего, например, слово *тетя* приходится искать в двух вариантах – *тетя* и *тётя*. По счастью, некоторые поисковые программы не делают разницы между Е и Ё и выдают по-разному оформленные слова в едином перечне ответов на запрос, но это имеет место лишь тогда, когда знак Ё идентифицируется данной программой как Е – а это отмечается не всегда.

Для решения данных проблем при пополнении корпуса и его редактировании нам приходится внимательно просматривать новый текст и заменять нечитаемые символы, проявляющиеся вместо Ё, или самое Ё знаком Е. Это выполнимо, однако составляет трудности на большом объеме текстов.

3. Слитно - раздельно

Слитные, дефисные и раздельные написания, в особенности первые две названные группы, довольно непротиворечиво выявляются в корпусе при поиске слов и словосочетаний, причем полученные наблюдения оказываются информативными: так, авторам удалось выявить более десятка написаний слова

подшофе в раздельном, дефисном и слитном оформлении, которые вполне корреспондируют нормам отдельных периодов истории русской орфографии и наглядно иллюстрируют их вариантность.

4. Стандартизация русской орфографии

Другая проблема, приобретающая особую важность в последнее время – это представление русских текстов в аутентичной графике и орфографии в целом или в графике и орфографии, действовавшей до 1917 г. включительно. Стремительно растет число таких текстов в Библиотеке Мошкова, являющейся ценным источником для текстового материала, в старой орфографии выставлены в Интернете, например, тексты В.И. Даля и Б.К. Зайцева. Наличие таких текстов очень ценно для исследования русской лексики.

Однако их появление сопряжено с целым клубком сложностей. Знаки дореволюционной гражданской кириллицы – ять, фита, ижица, даже если и отражаются аутентично в скопированных и интегрированных в корпус текстах, то все равно не распознаются поисковыми программами, имеющими собственные шрифтовые настройки. Буква *і* распознается программами, но возникают проблемы с идентификацией слов и форм со стороны морфологии: ни один известный морфологический анализатор не работает с русской морфологией в старой орфографии (это касается и форм с конечным ъ). Каждый очередной текст в старой русской орфографии побуждает к размышлениям: заменить отмененные в 1918 г. буквы и привести текст к современной орфографии или дождаться появления более совершенных программ, которые будут нивелировать различия между дореформенным и послереформенным написанием слов. Впрочем, тут встречаются подлинные филологические «шедевры» – например, тексты произведений В.И. Даля, где сохраняются ять, ъ, і, и т.п., но при этом оказывается расставленным Ё, отсутствующее в оригинальных текстах.

Только в наши дни становится понятным, что разработка русских шрифтов вполне могла бы базироваться не на современном русском алфавите, а на дореволюционном, в который, как известно, реформа 1918 года не добавила ничего, кроме апострофа вместо Ъ внутри слова. То есть во всех компьютерах мог бы использоваться комплект знаков, соответствующий современной русской графике, но те же шрифты позволяли бы читать и редактировать тексты в дореволюционной орфографии, а также большинство древнерусских текстов в упрощенной графике, равно как и преобразовывать тексты из одной формы русской орфографии в другую без потери шрифтовой идентичности. В настоящее время создание корпусов древнерусских текстов представляет собой сложную задачу именно в силу того, что отсутствует единый стандарт для представления русских текстов в разной графике, что представляется почти нереальным оперативно привести имеющиеся в достаточном количестве тексты к единому графическому облику.

Изучение орфографических вариантов слов, относящихся к периодам 1920–1950-х годов по корпусам текстов и по «Библиотеке лексикографа», безусловно, сопряжено с рядом сложностей. Во-первых, далеко не все тексты этого периода существуют в электронном виде в авторской орфографии: априори чаще всего в них представлена орфография последнего издания, хотя есть возможность вводить в корпус отдельные тексты, преобразованные из форматов PDF и Djvu. Во-вторых, для того, чтобы обнаруживать те или иные написания в корпусе, надо иметь их список, который пока в научном обороте отсутствует. Тем не менее, корпусы и здесь составляют альтернативу классической словарной картотеке, поскольку мы не располагаем данными, сохранялись ли в картотеках и в каком объеме авторские написания слов.

Так или иначе, исследование графико-орфографических вариантов слов при помощи корпусов текстов или «Библиотеки лексикографа» намного – на несколько порядков – увеличивает

объем доступного материала, хотя и оставляет желать много лучшего в отношении исходных данных. Впрочем, исторические словари русской орфографии как жанр в отечественной лексикографической традиции пока отсутствуют.