

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. М. Т. КАЛАШНИКОВА

Информационные технологии и письменное наследие

El'Manuscript-2012

Материалы IV международной научной конференции
Петрозаводск, 3–8 сентября 2012 года

Петрозаводск, Ижевск
2012

УДК 004.9
ББК 81.11+81.2-0
И741

Изданы при поддержке гранта РФФИ (проект № 12-06-06061-г),
гранта РГНФ (проект № 12-04-14154-г) и в рамках реализации комплекса
мероприятий Программы стратегического развития ПетрГУ на 2012-2016 г.

Ответственные редакторы:

В. А. Баранов, д-р филол. наук, проф.

А. Г. Варфоломеев, канд. физ.-мат. наук, доц.

Информационные технологии и письменное наследие [Текст] :
И741 материалы IV междунар. науч. конф. (Петрозаводск, 3–8 сентября
2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. — Петрозаводск ;
Ижевск, 2012. — 328 с.

ISBN 978-5-8021-1402-5

Сборник содержит материалы конференции, посвященной современ-
ным электронным средствам хранения, описания, обработки, исследования
и публикации памятников письменности и исторических источников.

УДК 004.9
ББК 81.11+81.2-0

ISBN 978-5-8021-1402-5

© Петрозаводский государственный
университет, 2012
© Ижевский государственный технический
университет им. М. Т. Калашникова, 2012

ЭЛЕКТРОННЫЙ СЛОВАРЬ ВАРИАНТОВ НА ОСНОВЕ ТЕКСТОВ XVIII века¹

С. О. Савчук

Институт русского языка им. В.В. Виноградова РАН, Москва

Electronic dictionary of variants is being created on the basis of the texts of the 18th century in the Russian National corpus. For its formation the list of the unidentified word-forms supplied with hypothetic lemmas is used. The database of these word-forms gives essentially valuable material for the analysis of the orthographic, phonetic and different kinds of grammar variation in the 18th-century language. At the same time the study of variation has practical use for customizing tagger according to special cases which are typical for 18th-century texts.

Электронный словарь вариантов создается на базе текстов XVIII в. в составе Национального корпуса русского языка. На первом этапе словарь будет использован для формирования словника разрабатываемого морфологического анализатора для текстов XVIII–XIX в. В дальнейшем возможно создание электронного ресурса, который пополнит семейство электронных словарей на основе НКРЯ [Словари]. Как известно, тексты XVIII в. отличаются крайне высокой степенью вариативности, которая слабо отражена в существующих словарях и практически не описана в нормативных грамматиках. Все это делает невозможной корректную работу морфологического анализатора, с помощью которого выполняется автоматическая аннотация текстов, поскольку в словаре анализатора вариативность также не учтена. Так, например, слово *фельдмаршал* присутствует в словаре в единственном орфографическом облике, в то время как в текстах петровского периода оно может быть передано самыми разными способами: *фелтмаршал*, *Фельтмаршал*, *фельтьмаршал*, *фелтмаршел*, *Фелт Маршал*, *Фельт Маршал* и др., — ни один из которых не опознается как орфографический вариант одной и той же леммы.

Учитывая особенности текстов XVIII–XIX в., формирование словаря морфологического анализатора для диахронического модуля НКРЯ решено осуществлять по двум направлениям. Первое направление предполагает расширение словника за счет включения в него материалов исторических словарей (САР, СЦСРЯ и др.). Второе направление сосредоточено на пополнении словника словами, извлеченными непосредственно из корпуса текстов XVIII–XIX вв., которые не учтены в лингвистических источниках. Работа в рамках второго направления проводилась в несколько этапов.

1. Сформирован экспериментальный корпус текстов XVIII – начала XIX в. общим объемом около 4 млн словоупотреблений. Проведена обработка текстового массива, получены частотные списки словоформ, содержащие сведения об их грамматических характеристиках.

2. Построен частотный словарь словоформ, не получающих предсказанного разбора. Поскольку соответствующие им лексемы отсутствуют в словаре морфологического анализатора, эти словоформы снабжены гипотетическими разборами, правильными или ошибочными.

¹ Работа выполнена при поддержке: Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика»; РФФИ (грант 10–06–00151а).

3. Организована база данных несловарных словоформ, имеющая следующие поля: словоформа, сгенерированная лемма, отсылочная (нормализованная) лемма, грамматические признаки леммы, грамматические признаки словоформы, тип варианта, сведения об авторе и дате создания текста, в котором зафиксирована форма.

4. Проведен анализ 4000 единиц базы данных: для каждой словоформы приписана или выбрана из предложенного списка правильная лемма, приписаны отсылочная (нормализованная) словоформа и нормализованная лемма, грамматическая информация проверена по корпусу и определены правильные грамматические характеристики.

Поскольку анализ базы выполняется в ручном режиме, необходимо было оценить целесообразность этой трудоемкой работы, эффективность которой выражалась бы в повышении качества разбора текстов. Проведенный анализ показал, что исправление ошибочных разборов почти вдвое сокращает количество порожденных гипотез, что является достаточно высоким показателем. По частеречной принадлежности проанализированные единицы распределились следующим образом. Имена существительные составляют более половины всех словоформ, из них имена собственные (личные имена, фамилии, топонимы) — около 51%. Прилагательные составляют около 40%, больше половины из них приходится на долю прилагательных, образованных от собственных имен. Оставшиеся 9% словоформ представляют собой формы глаголов, наречия и служебные части речи. Таким образом, в целом около половины списка форм с непредсказуемым разбором составляют имена собственные и образованные от них прилагательные. В остальной части списка заметное место занимают лексемы, отсутствующие в словаре (*ажуатёр, анатомить, ажно, аще, аркебузировать, бомбаст, безнуждно* и пр.), морфологические архаизмы (*бранех, гресех, духови, острови, знаеши, бяше, бяху, являшеся* и пр.) и варианты различного рода.

5. Следующий этап работы связан с формированием базы данных вариантов. Была проведена разметка вариантов, полученных на основании анализа текстов XVIII в., выделены орфографические, фонетические и морфологические варианты, не учтённые в грамматическом словаре НКРЯ. Среди них самую заметную часть составляют орфографические варианты — около 76% (*армея, болиший, брегодир, ариергард / арриергард, будтобы*), которые бывает трудно характеризовать отдельно от фонетических (*англинский, английнский, аглинский, амврозия, амфилада, анбиции, армонический, аржаной, арженой, ариветика, бонбардирование, гарнадеры* и под.). Морфологические варианты составляют около 2% (*армяна, армянов* вм. *армяне, армян, Апеннинов* вм. *Апеннин, знамены* вм. *знамена, две редуты* вм. *два редута* и др.); словообразовательные — 18% (*Аничков, Аничковский, Аничковый, аравитский, антический, артилерный, архивский, башикирцов, бедство, бездельство* и под.).

6. Проведена оценка единиц базы данных с целью отбора кандидатов на пополнение словаря морфологического анализатора. Основным критерием является частотность словоформы. Так, вряд ли необходимо включать в электронный словарь корпуса низкочастотные имена собственные (такие как *Азбад, Айшедуд, д'Аркур, Ардильер, Блюмфельберг* и под.), однако высокочастотные в текстах XVIII–XIX в. слова должны в нем присутствовать (*аншеф, благой, вышеписанный, вышепомянутый, государство, доношение, деташемент, ретирада, апробовать, глаголить, испросить, аще, егда* и под.). Другим немаловажным критерием является формальный облик слова: если словоформа не дает возможности более или менее точно предсказать его лемму и грамматические характеристики, слово с этой информацией следует поместить в словарь. В особенности это касается архаизмов. Так для формы *агнчий* предлагается 12 разборов, из которых только 2 правильных, для *ангельстии* — 16 вместо 1, для *алкаличных* — 11 вместо 3, для *англяне* — 24 разбора вместо 1 правильного, для *аллилуия, алчушу* — 11 вместо 1, и т. д. Наличие вариантов также является основанием для помещения слова в словарь, а если какой-либо вариант уже присутствует в словаре, то новый приписывается с помощью соответствующих ссылок.

Завершение формирования базы данных и внедрение ее в словарь позволит проверить гипотезу о том, что расширение словаря морфологического анализатора и настройки его на определенные тексты способно повысить качество автоматической морфологической разметки как текстов XVIII в., так и всего диахронического модуля в целом. Электронный словарь вариантов, созданный на основе НКРЯ, будет полезен для изучения особенностей языка XVIII в. и истории становления литературной нормы.

Список литературы

Словари — Словари, созданные на основе Национального корпуса русского языка [сайт]. [2008]. URL: <http://dict.ruslang.ru> (дата обращения: 15.04.2012).