Корпус текстов XVIII века в составе Национального корпуса русского языка: проблемы и перспективы[1]

С. О. Савчук

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

Формирование подкорпуса текстов XVIII века ведется с 2006 г. в рамках сотрудничества Казанского университета и Института русского языка им. В. В. Виноградова РАН. В 2006 г. был создан пилотный корпус [Савчук и др. 2006], к настоящему времени его объем увеличен до 2 млн. словоупотреблений за счет расширения круга авторов и жанрового состава текстов. В корпусе пропорционально представлены различные сферы функционирования языка: публицистика составляет 24 %, учебно-научные тексты – 17 %, официально-деловые – 11 %, церковно-богословские – 19 %, бытовые (личная переписка, дневники) – 5 %, художественная проза – 24 %. Поэзия XVIII в. (более 0,5 млн. словоупотреблений) представлена в поэтическом корпусе.

Основная задача, которая ставилась на первом этапе создания пилотного корпуса, заключалась в том, чтобы проверить возможность обработки и описания текстов, принадлежащих прошлым состояниям языка, с помощью средств, разработанных для аннотации современных текстов, с целью выявления способности системы разметки адаптироваться к новому лингвистическому материалу. Эта задача была успешно решена, доказательством чему служит функционирующий корпус и исследования, выполняемые на его основе [Савчук и др. 2006; Savchuk 2007].

Задачей второго этапа стал анализ проблем, возникших при формировании корпуса, с целью оптимизации процесса его создания и использования.

1. *Проблема выбора источников текстов.* Ресурсы электронных филологических библиотек [РВБ, ФЭБ, ImVerden], отличающиеся высокой культурой подготовки текстов и в первую очередь привлекавшиеся для формирования пилотного корпуса, оказались практически исчерпанными. Электронные версии из исторических и юридических библиотек [Библиотека,

[1] Работа выполнена при поддержке РГНФ, грант № 06-04-03817в.

Военная, Восточная, Хронос и др.], к сожалению, часто не отвечают стандартам качества подготовки текстов, установленным для корпуса, и нуждаются в серьезном редактировании. В связи с этим приходится искать источники в электронных библиотеках, хранящих книги в графических форматах или самим заниматься оцифровкой типографских изданий.

2. *Проблема редактирования текстов и орфографической унификации.* При подготовке текстов XVIII-XIX вв. для включения в корпус избрана стратегия, не требующая масштабной унификации: электронная версия должна в общем соответствовать печатной. Поэтому если воспроизводится современное издание текстов XVIII в., то орфография в нем будет в основном соответствовать правилам 1956 года, при воспроизведении дореволюционного издания в нем сохраняются все особенности орфографических норм соответствующего периода, за исключением тех изменений в графике, которые были внесены реформой 1918 года. Это означает, что перед включением в корпус тексты должны пройти умеренную орфографическую модификацию, в результате которой в них устраняются элементы, отмененные реформой 1918 г. (например, буквы *ѣ, ѳ, ѳ*; буква **ъ** после твердого согласного в конце слова, *і* перед гласным и *й* и т. д.).

3. *Проблема лингвистической аннотации.* Морфологическая разметка, в процессе которой выделяются словоформы и каждой словоформе приписывается информация о ее лексемной принадлежности и о совокупности ее грамматических признаков, производится на основной части корпуса в автоматическом режиме с помощью специальных программ-парсеров, использующих встроенные морфологические словари. Программа порождает все возможные разборы словоформы, а в случае отсутствия словоформы в словаре строит гипотезы относительно ее лексемной принадлежности и предлагает гипотетические разборы [Ляшевская и др. 2005: 117]. Задача оптимизации поиска состоит прежде всего в уменьшении количества маловероятных разборов.

Предварительный анализ вхождений несловарных форм показал, что около 45 % из них представляют лексемы, отсутствующие в словаре, используемом анализатором (архаизмы, собственные имена и производные

от них), среди них есть весьма частотные: *тако* (297), *всяко* (101), *Плиний* (111), *васильевском* (71), *порфирогенит* (26) и др. Больше половины контекстов с несловарными формами выявляют различные варианты входящих в словарь слов, из них более 20 % составляют орфографические варианты (*полаты*, *толко*, *протчих*, *зделать, домогатца*), около 17 % – морфологические (*егеров* ср. норм. *егерей*; *турков* ср. *турок*; *клянуся* ср. *клянусь*; *по сту* ср. *по сто*), около 14 % – словообразовательные (*авангардия* ср. *авангард*; *канцелярный* ср. *канцелярский*; *самодержавство* ср. *самодержавие*), около 3 % – фонетические (*гистория*, *эсквадра*, *гранодеры*, *провинциял-фискал*).

Таким образом, практика создания корпуса XVIII в. подтверждает, что проблема совершенствования морфологической разметки текстов с большим количеством нестандартных форм является общей для всех текстов, язык которых выходит за пределы *современной письменной литературной нормы*. Решение этой проблемы следует искать по крайней мере в трех направлениях: 1) в нормализации орфографии, 2) в пополнении словаря корпуса, 3) в обучении программ-парсеров на специфическом для каждого корпуса текстовом материале.

Для каждого корпуса, по-видимому, должна избираться наиболее оптимальная тактика работы, учитывающая степень вариативности текстов и особенности состава несловарных единиц. В частности, для корпуса XVIII в., характеризующегося высокой степенью орфографической вариативности, эффективна орфографическая нормализация на этапе технического редактирования и структурной разметки текстов. При таком способе каждому ненормативному написанию приписывается нормативная форма: *естли*{если*}, *зделать*{сделать*}, *доволно*{довольно*} и др. В процессе морфологической разметки разбирается нормативная форма, а набор грамматических признаков приписывается всему комплексу, так что при лексико-грамматическом поиске в корпусе выдаются контексты, содержащие это слово во всех вариантах написания.

На ближайшее будущее разработчики корпуса текстов XVIII в. ставят перед собой следующие задачи: во-первых, пополнение корпуса новыми

текстами (прежде всего первой трети XVIII в.), подготовка и включение в состав корпуса редких текстов (частных писем, деловой переписки, газет, старопечатных книг), прошедших процесс соответствующей орфографической обработки; во-вторых, полный анализ несловарных форм, выделенных в текстах XVIII в. (всего около 3000 словоформ), и пополнение ими словаря корпуса.

В качестве задачи на отдаленную перспективу можно было бы рассматривать создание комплексного информационного ресурса, объединяющего электронную библиотеку, корпус текстов в дореволюционной орфографии [Соловьев и др. 2006] и корпус текстов в современной орфографии. Такой ресурс мог бы удовлетворить интересы специалистов разных профилей, изучающих культурное наследие XVIII века.

Список литературы

Библиотека – *Библиотека электронных ресурсов Исторического факультета МГУ им. М.В.Ломоносова* [Электронный ресурс]. – Режим доступа : http://www.hist.msu.ru/ER/index.html, свободный. – Загл. с титул. страницы.

Военная – *Военная литература* [Электронный ресурс]. – Режим доступа : http://militera.lib.ru/, свободный. – Загл. с титул. страницы.

Восточная – *Восточная литература* [Электронный ресурс]. – Режим доступа : http://www.vostlit.info/haupt-Dateien/index-Dateien/H.phtml, свободный. – Загл. с титул. страницы.

Ляшевская и др. 2005 – *Ляшевская, О. Н*. О морфологическом стандарте Национального корпуса русского языка / О. Н. Ляшевская, В. А. Плунгян, Д. В. Сичинава // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. – М.: Индрик, 2005. – С. 111-134.

НКРЯ – *Национальный корпус русского языка* [Электронный ресурс]. – Режим доступа : www.ruscorpora.ru, свободный. – Загл. с титул. страницы.

РВБ – *Российская виртуальная библиотека* [Электронный ресурс]. – Режим доступа : http:// www.rvb.ru, свободный. – Загл. с титул. страницы.

Савчук и др. 2006 – *Савчук, С. О.* Подкорпус текстов XVIII века в составе Национального корпуса русского языка: из опыта работы / С. О. Савчук,

Д. В. Сичинава, И. И. Гарипов. – Режим доступа : http://fccl.ksu.ru/issue_spec/docs/Savchuk_Sichinava_Garipov.doc, свободный. – Загл. с титул. страницы.

Соловьев и др. 2006 – *Соловьев, В. Д.* Корпус русского языка XVIII века: текущее состояние/ В. Д. Соловьев, Р. Б. Ахтямов // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам : материалы междунар. научн. конф., Ижевск, 13-17 июля 2006 г. – Ижевск, 2006. – С. 156-160.

ФЭБ – *Фундаментальная электронная библиотека «Русская литература и фольклор»* [Электронный ресурс]. – Режим доступа : http://www.feb-web.ru/, свободный. – Загл. с титул. страницы.

Хронос – *ХРОНОС* [Электронный ресурс]. – Режим доступа : http://hronos.km.ru, свободный. – Загл. с титул. страницы.

ImVerden – *ImVerden* [Электронный ресурс]. – Режим доступа : http://www.imverden.de, свободный. – Загл. с титул. страницы.

Savchuk 2007 – *Savchuk*, *Svetlana*. Corpus-based Investigation of Language Change: the Case of RNC // Proceedings of the Corpus Linguistics Conference CL2007 University of Birmingham, UK, 27-30 July 2007 / Matthew Davies, Paul Rayson, Susan Hunston, Pernilla Danielsson (eds.). – Режим доступа : http://ucrel.lancs.ac.uk/publications/CL2007/final/181/181_Paper.pdf, свободный. – Загл. с титул. страницы.

The corpus of 18th-century texts as a part of the Russian National Corpus: problems and prospects

Svetlana O. Savchuk

Vinogradov Institute of the Russian Language of the Russian Academy of Sciences, Moscow, Russia

The paper presents the results of a project aimed at the creation of a sub-corpus of 18th-century Russian texts within the RNC. The main problems concerned with text preparation and morphological annotation and prospects for Corpus development are discussed.