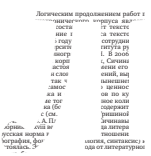
С. О. Савчук,
Д. В. Сичинава

Корпус русских текстов XVIII века

в составе Национального корпуса русского языка: проблемы и перспективы[1]



огическим продолжением работ по созданию диахронического корпуса является расширение его состава за счет текстов XVIII века. Формирование подкорпуса текстов XVIII века начато в 2006 году в рамках сотрудничества Казанского университета и Института русского языка им. В. В. Виноградова РАН.

В 2006 г. был создан пилотный корпус [Савчук, Сичинава, Гарипов 2006], к настоящему времени его объем увеличен до 2 млн словоупотреблений, выровнен состав текстов, так что уже в нынешнем виде корпус имеет самостоятельную ценность для историков языка и специалистов по культуре XVIII века. Кроме того, существенное количество текстов XVIII века (более 438 тыс.) содержит поэтический корпус (см. статью Е. А. Гришиной, К. М. Корчагина, В. А. Плунгяна и Д. В. Сичинавы в наст. сборнике).

XVIII век — период, когда литературная русская норма в самых разных отношениях (орфография, фонетика, морфология, синтаксис) не устоялась. Это период перехода от литературного

[1] Работа выполнена при поддержке РГНФ, грант № 06-04-03817в и № 07–04–12147в («Большой корпус русского языка XVIII в.»)

языка, базирующегося на церковнославянском, к языку нового типа, так или иначе отражающему собственно русскую языковую систему. История русского литературного языка XVIII века пока разработана несколько меньше (по крайней мере, с чисто лингвистической точки зрения), чем языка допетровского времени или следующего периода — языка XIX в. (следует назвать монографии Живов 1996, Живов 2004, Успенский 1985). Исследование *литературного языка* иногда, к сожалению, подменяется исследованием *языка литературы* — нескольких крупнейших писателей. А ведь особенные линии эволюции определяют нормы различных жанров этой эпохи: язык официально-деловых документов, публицистики, проповедей, частной переписки и проч. Корпус, включающий в себя тексты самых разных жанров, призван облегчить будущим исследователям задачу разностороннего исследования языка XVIII века.

В существующих работах по истории русского литературного языка принято выделять два [Горшков 1969] или три периода [Виноградов 1978, Винокур 1959], связанных с XVIII веком:

1) Петровское время (конец XVII — первая треть XVIII в.)—период «смешения и объединения—несколько механического—живой разговорной речи, славянизмов и европеизмов на основе государственно-делового языка» и формирования новых стилей «гражданского посредственного наречия» и литературных стилей, занимающих «промежуточное положение между возвышенным славянским слогом и простой разговорной речью».

2) Ломоносовский период (40–50-е гг.—конец XVIII в.)—период стилистической регламентации и нормализации нового русского литературного языка на основе теории трех стилей.

3) Карамзинский период (конец XVIII—начало XIX в.)—реорганизация литературного языка, выразившаяся в отмене жанровых ограничений, в создании «нового слога российского языка»—средней литературной нормы, близкой к разговорному языку образованного общества [Виноградов 1978].

В пилотный корпус текстов XVIII века включены прозаические тексты, относящиеся в основном ко второму и третьему периоду и представляющие все сферы функционирования языка в разнообразии жанровых разновидностей.

Художественная сфера представлена прозаическими произведениями писателей, оказавших заметное влияние на процесс формирования литературного языка: Н. М. Карамзин, И. А. Крылов, Н. И. Новиков, А. А. Нартов, А. Н. Радищев, Д. И. Фонвизин, М. Д. Чулков. Стихотворные тексты 14 авторов (И. Ф. Богдановича, И. С. Баркова, Г. Р. Державина, И. И. Дмитриева, А. Д. Кантемира, И. А. Крылова, М. В. Ломоносова, А. П. Сумарокова, В. К. Тредиаковского, И. И. Хемницера, М. М. Хераскова и др.) входят в состав поэтического корпуса.

Сфера публицистики представлена прежде всего сатирическими статьями Н. И. Новикова в журналах «Трутень», «Пустомеля», «Кошелек», «Живописец», полемикой Н. И. Новикова с Екатериной II, статьями и рецензиями И. А. Крылова, статьями и очерками на общественно-политические темы Д. И. Фонвизина, А. Н. Радищева, философским трактатом Г. Сковороды, памфлетом М. М. Щербатова, мемуарами А. Т. Болотова, П. А. Левашова, Я. П. Шаховского.

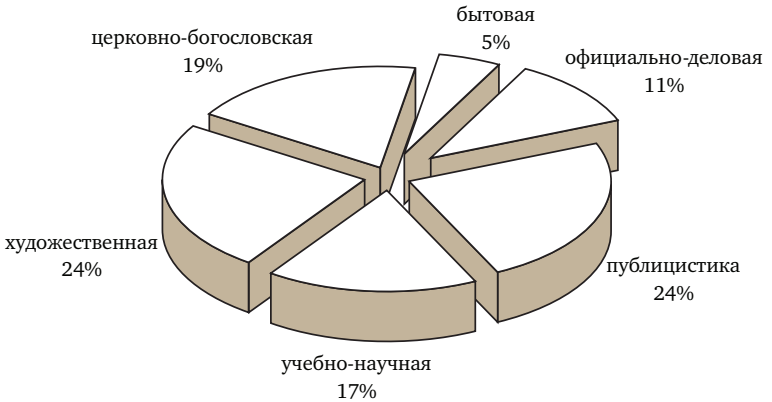
Учебно-научная сфера отражена в сочинениях А. Н. Радищева из области экономики, права, истории, политики, в филологических сочинениях М. В. Ломоносова, Д. И. Фонвизина, Н. И. Новикова, в трудах историка В. Н. Татищева. Представлены научные трактаты, статьи, рецензии, инструкции, словари.

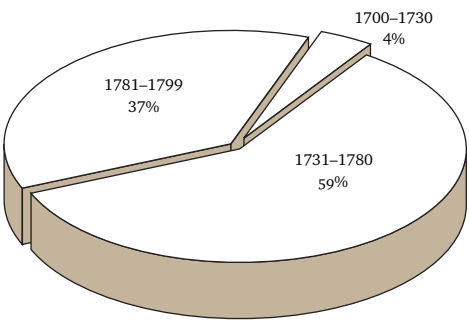
Официально-деловая сфера представлена жанрами законодательных, правовых, дипломатических и деловых документов (указ, манифест, воинский устав, проект, приказ, дипломатический договор, служебная записка, военное донесение, прошение, завещание). Это прежде всего документы эпохи Петра I, Екатерины II.

Бытовая сфера — это личные письма Н. М. Карамзина, А. Н. Радищева, Д. И. Фонвизина, И. Ф. Богдановича, А. А. Боратынского (отца поэта), Н. А. Львова, Г. Сковороды, А. В. Суворова, дневники С. А. Порошина.

Церковно-богословская сфера представлена в сочинениях Платона (Левшина), Архиепископа Московского и Калужского, блестящего представителя духовного красноречия, и Феофана (Прокоповича). Среди жанров—слово, поучение, катехизис, краткий учебник по закону Божию.

Количественное распределение текстов по основным сферам функционирования и периодам создания представлено на диаграммах.





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 гибкости системы разметки и ее адаптивности к новому лингвистическому материалу. Эта задача была успешно решена, доказательством чему служит функционирующий корпус и исследования, выполненные на его основе [Савчук 2006; Савчук, Гришина 2008].

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

Эту проблему приходится решать как создателям электронных библиотек, так и разработчикам корпусов. Однако в отличие от электронных библиотек, в которых можно разместить несколько вариантов/ редакций одного и того же текста (как это делается, например, в ФЭБе или в РВБ)[2], корпус включает единственную электронную версию, в связи с чем более остро стоит вопрос выбора источника и качества его редактирования.

Необходимо разграничивать три типа источников.

1) Первичные источники—старопечатные книги и рукописные тексты, которые для введения в состав корпуса проходят полный цикл подготовки, включающий оцифровку, распознавание, корректуру и редактирование электронной версии.

2) Печатные издания (как дореволюционные, так и современные), электронные версии которых изготавливаются для корпуса. Цикл подготовки таких текстов отличается от первого случая тем, что приходится оценивать качество издания с точки зрения соответствия оригиналу и, если оно не единственное, выбирать наиболее авторитетное.

3) Электронные версии текстов, взятые из электронных библиотек. В данном случае процесс подготовки значительно упрощается и сводится к корректуре — сверке электронной версии с первоисточником или, в случае его недоступности, с авторитетным изданием и структурной разметке и редактированию электронной версии.

Ресурсы электронных филологических библиотек (РВБ, ФЭБ, ImWerden), отличающиеся высокой культурой подготовки текстов и в первую очередь привлекавшиеся для формирования пилотного корпуса, оказались к настоящему времени практически исчерпанными. Электронные версии из исторических и юридических библиотек (Библиотека, 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] Об эдиционных принципах филологических электронных библиотек см., например, *http://www.rvb.ru/about/principles.html, http://feb-web.ru/feb/feb/about1.htm#Lo4*

2. Проблема редактирования текстов и орфографической унификации

Специфика подкорпуса XVIII века (а также XIX-го и I-й половины XX-го) как части Национального корпуса состоит в том, что тексты, включенные в него, должны быть переданы только средствами современной орфографии, поскольку она лежит в основе всех средств грамматической разметки и поиска. В XX веке русская орфография дважды подвергалась реформированию: реформа 1918 года изменила графику и унифицировала ряд написаний (окончания прилагательных, причастий, местоимений, приставки на *–з* и др.), реформой 1956 года были отрегулированы написания отдельных категорий слов и морфем. Поэтому проблема редактирования оригинала, связанного с орфографической модернизацией текстов дореволюционных изданий, для корпуса XVIII века стоит очень остро. При этом каждый тип источников требует особого подхода.

При подготовке источников первого типа в НКРЯ приняты эдиционные принципы, общие для изданий академического типа или близких к ним, а также филологических электронных библиотек (например, РВБ). Орфография оригинала подвергается умеренной модернизации — модернизируются только такие написания, которые могут быть восстановлены автоматически (например, *ъ* после твердого согласного в конце слова, *і* перед гласным и *й*; замена *ѣ* на *е* и т. д.). Особенности орфографии первоисточника, не отрегулированные реформой 1918 года, сохраняются (*фелтьмаршал, салдаты, торелка* и под.).

При подготовке источников второго типа составители НКРЯ придерживаются основного общего принципа: электронная версия должна соответствовать печатной. Однако если текст издавался несколько раз, отдельные издания могут сильно отличаться друг от друга. Для текстов XVIII века эта проблема особенно актуальна, поскольку строгих правил, регламентирующих написание, в XVIII веке не существовало. Поэтому при последующих изданиях этих текстов они, как правило, подвергались редактированию с позиций действующих в момент публикации орфографических норм и правил. В отдельных случаях, когда текст освоен культурой и продолжает переиздаваться (и даже входит в школьную программу), этот процесс модернизации орфографии источника заходит очень далеко, так что, например, современные школьные издания повестей Н. М. Ка-

рамзина, басен И. А. Крылова, пьес Д. И. Фонвизина полностью соответствуют действующим с 1956 года правилам орфографии. Сравним фрагмент текста «Юности честное зерцало», представленный в «Хрестоматии по русской литературе XVIII века» (М.: Просвещение, 1979) и в издании XVIII века[3].

> Когда им говорить с *людьми*, то должно им благочинно, учтиво, вежливо, разумно, а не много говорить; потом слушать и других речи *не перебивать*, но дать все выговорить и *потом* мнение свое, что достойно, предъявить. Ежели случится дело и речь *печальная*, то надлежит при таких быть *печальну* и иметь сожаление. В радостном случае быть радостну и являть себе весела с веселыми.
>
> А в прямом деле и в постоянном быть постоянну, и других *людей рассудков* отнюдь не презирать и не отметать, но *ежели* чие мнение достойно и годно, то похвалять и в том *соглашаться; ежели* же которое *сумнительно*, в том себя оговорить, что в том ему *рассуждать* не достойно. А *ежели* в чем оспорить можно, то учинить с учтивостию и вежливыми словами, и дать свое *рассуждение* на то, *для чего*. А ежели кто совету пожелает или что поверит, то надлежит советовать *сколько* можно и поверенное дело содержать тайно.

А вот как этот фрагмент выглядит в оригинале (курсивом отмечены орфографические расхождения между двумя фрагментами).

> 7. Когда имъ говорїть с *людми*, то должно имъ благочїнно, учтїво, вѣжлїво, разумно, а не много говорїть. потом слушать, и другїхъ рѣчи *неперебївать*, но дать все выговорїть и *по томъ* мнѣнїе свое, что достоїно, предъявїть. Ежели случїтся дѣло и рѣчь *печалная*, то надлежїтъ при такїх быть *печалну,* и имѣть сожалѣнїе. въ радостномъ случае быть радостну, и являть себе весела съ веселыми.
>
> А въ прямомъ дѣлѣ и въ постоянномъ, быть постоянну, и другїхъ *людеи разсудковъ* отнюдь не презїрать и не отмѣтать. но *еже ли* чїе мнение достоїно и годно, то похвалять и въ томъ *соглашатца. еже ли* же которое *сумнїтелно*, въ томъ себя оговорїть, что въ томъ ему *разсуждать* не достоїно. А *еже ли* въ чемъ оспорїть можно, то учїнїть съ учтївостїю и вѣжлївыми словами, и дать свое *разсужденїе* на то, *длячего.* А ежели кто совѣту пожелаетъ или что поверїтъ, то надлежїтъ совѣтовать *сколко* можно и повѣренное дѣло содержать таїно.

---

[3] Юности честное зерцало или показанїе къ житеїскому обхождению. Собранное отъ разныхъ авторовъ. Напечатася повелѣнїемъ царскаго велїчества. В Санктпїтербурхѣ лѣта господня 1717 февраля 4 дня. — Факсимильное издание. М., 1976 (*http://elibrary.karelia.ru*)

Как видим, отличия между двумя версиями текста значительны: в учебном издании произведены не только графические замены (ѣ на е, ї на и или й, Ъ на конце слов), но и в соответствии с современными орфографическими нормами унифицированы отдельные написания: буквы Ь для обозначения мягкости согласных в середине слова (*людми — людьми, печалну — печальну, сколко — сколько*), приставки раз-/рас- (*разсудков — рассудков, разсуждение — рассуждение*), окончаний глаголов (*соглашатца — соглашаться*), слитного или раздельного написания предлогов, частиц (*еже ли — ежели, длячего — для чего*) и т.д.

Поэтому при подготовке электронных версий опубликованных текстов большое внимание уделяется выбору авторитетного издания, и в дальнейшем электронная версия приводится в соответствие с печатным оригиналом: если воспроизводится современное издание текстов XVIII века, то орфография в нем будет 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.

Наконец, третий тип источников — тексты из электронных библиотек — требует оценки качества электронных версий и их соответствия оригиналу. Как показала практика, качество электронных версий, взятых из филологических библиотек (ФЭБ, РВБ, ImWerden) таково, что обычно не требует дополнительной корректуры, и предварительная подготовка текста для включения в корпус сводится к техническому редактированию и структурной разметке текста. Электронные версии из исторических и юридических библиотек нуждаются в дополнительном редактировании и текстологической подготовке, поскольку тексты могут быть представлены в отрывках, с купюрами, в орфографии, модернизация которой проведена непоследовательно.

Приведем в качестве примера результат сравнения орфографии небольшого фрагмента «Военного устава 1716 года (Раздел 3. Краткое изображение процессов или судебных тяжеб)» из двух электронных библиотек.

Орфограмма	**1. Военно-исторический проект «Адъютант!»**	**2. Хрестоматия по истории государства и права России / Ю. П. Титов. —М., 2002**
Окончания прил., прич., мест. Р.ед. м-ср. -аго, -яго И.,В. ж. мн. -ыя, -ія	достойнаго некотораго высокаго разныя происходящия последующия государственныя целаго происходящия касающияся другаго которыя прочия	достойнаго некоторого высокого разные происходящие последующие государственные целаго происходящыя касающыяся другаго которые протчие
Приставки из-, воз-, раз-, роз-, низ-, без-, через-, чрез-	разъискиваются разделяется разсуждаем	разыскиваются розделяется разсуждаем
Слитно/раздельно/ через дефис	притом	при том
Двойные согласные	процессах	процесах
Мягкость согласных	обстоятельства начальства генеральной генеральном	обстоятельства началства генералной генералном
Прочие орфограммы в корне	между между между Фельдмаршала причины причины прочия	междо между междо фелтьмаршала притчины притчины протчие
Орфограммы в аффиксах	находятся	находятца
Прописная/строчная	Офицеров Фельдмаршала	офицеров фелтьмаршала

Первая электронная версия, опубликованная на сайте *http://adjudant.ru*, восходит к изданию XVIII в.: «Военной устав с Артикулом военным, при котором приложены толкования, также

с кратким содержанием процессов, экзерциciею, церемониями, и должностьми полковых чинов». Вторым тиснением напечатан в Санктпетербурге. При Императорской Академии Наук 1748 года». Модернизация орфографии произведена создателями сайта: «В интернет-версии по большей части сохранена орфография книги-источника. Для удобства чтения заменено написание отдельных слов в соответствии с современными правилами (например, *потомуж—потому ж, отом—о том,* и т.п.). В некоторых частях заменены окончания (*великаго—великого, оной—оный*)». Вторая версия изготовлена по современному учебному изданию: Хрестоматия по истории государства и права России / Ю. П. Титов (М., 2002), следовательно, унификация орфографии — дело рук автора-составителя и редакторов издания.

Можно заметить, что в обоих изданиях модернизация орфографии проведена непоследовательно: непонятны принципы, по которым публикаторы в одних случаях предпочитают современный вариант написания, а в других—дореформенный (например, в первой версии избирается современный способ обозначения мягкости согласных внутри слова, написания отдельных корней, глаголов на *-ся*, но архаичный способ написания окончаний прилагательных, причастий, местоимений, отмененный реформой 1918 года). В целом электронная публикация на сайте «Адъютант!» кажется более привлекательной хотя бы потому, что в ней меньше внутритекстовых несоответствий, которыми изобилует второе издание (ср. *достойнаго, другаго* и *высокого, некоторого*; *происходящие, государственные* и *происходящыя, касающыяся*, *междо* и *между*, *обстоятельства* и *началства*).

Однако модернизация графики и орфографии еще не снимает проблему орфографических вариантов, которая может быть решена путем нормализации орфографии и будет рассмотрена в связи с общей проблемой вариативности.

3. Проблема лингвистической аннотации

Другая важная проблема, которую приходится решать в связи с созданием корпуса текстов XVIII в., является специфически корпусной и связана с лингвистической аннотацией. 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, Плунгян, Сичинава 2006: 117].

Гипотезы относительно грамматических характеристик отсутствующих в словаре словоформ (в разборах они имеют помету bastard) могут быть правильными; вероятность правильных разборов особенно высока в случае присутствия в составе этих словоформ современных аффиксов, например:

самодержавству
обосурманился
Гистория

Однако чаще порождаемые программой гипотетические разборы являются ошибочными, что создает большое количество шума при поиске:

 фортеции
тако

поехал *одоль* по правую сторону

<span title="одоливать = 2p,act,bastard,imper,pf,sg,v; одоля = bastard,f,gen,inan,pl,s| bastard,f,gen,inan,pl,s; одоль = acc,bastard,f, inan,s,sg|acc,bastard,inan,m,s,sg|adv,bastard| bastard,f,inan,nom,s,sg|bastard,inan,m,nom,s,sg">одоль

уже много тех *эксемпелев* (образов) есть

эксемпелев

Анализ грамматических разборов показал, что количество несловарных словоформ в текстах XVIII в. превышает показатели, характерные для письменных текстов, однако в сравнении с диалектными текстами и текстами электронной коммуникации, как видно из таблицы, эти различия невелики.

Подкорпус	**Объем подкорпуса**	**Количество несловарных словоформ**	**Соотношение в %**
XVIII	1106403	56695	5,1%
XIX	23730265	7009531	2,9%
XX-1	25902512	2834806	3,2%
XX-2-публиц	40440252	1390433	3,4%
XX-2-худож	35065938	747032	2,1%
XX-2-разг	4382391	71644	1,6%
XX-2-электр	1192121	83408	6,9%
XX-2-диал	138961	9045	6,5%

Предварительный анализ вхождений несловарных форм обнаружил, что около 45% из них представляют собственно новые лексемы, не включенные в словарь корпуса (архаизмы, историзмы, собственные имена и производные от них), среди них весьма частотные; особо надо выделить наречия образа действия на *–ко*, из которых первые два можно толковать как морфологические варианты современных наречий: *тако* (297), *всяко* (101), *инако* (92); из имён собственных, например—*Плиний* (111), *Васильевском* (71). Характерны целые архаичные модели словообразования, например, церковнославянские по происхождению слова на *благо-* (отмечены 22 таких слова, не предусмотренные современными словарями, например, *благополезный, благоутробно, благогласие*) или продуктивная отрицательная модель на *без-* (*безженство, безместный* и особо замечательное по семантике *безотрицательно*).

Больше половины контекстов с несловарными формами выявляют различные варианты входящих в словарь слов — орфографические (более 20%), морфологические (около 17%), словообразовательные (14%), фонетические (около 3%).

К частотным орфографическим вариантам относятся: *полаты* (56), *толко* (77), *одново* (3), *ево* (92), *лутче* (21), *протчих* (21), *протчим* (15), *естли* (10), *однакож* (55), *зделать* (27), *денги* (14),

возмет (6), *комисар* (9), *коммисия* (3), *домогатца* (3), *явятца*(3), *чинитца* (4) и др. Особенно они свойственны нередким для XVIII в. текстам со «свободной» орфографической установкой, например, в частной переписке или в отдельных публикациях вроде «Письма к другу, жительствующему в Тобольске» А. Н. Радищева.

Морфологические варианты представляют собой формы слов (как входящих, так и не входящих в словарь корпуса), которые не соответствуют морфологическим нормам современного русского языка (но могут быть употребительны в современном просторечии, диалектах и т.д.): *совестию*, *приязнию* (ср. *совестью*, *приязнью*), *клянуся*, *боялися* (ср. *клянусь*, *боялись*), *хощу* с церковнославянским чередованием (ср. *хочу*), *произвесть* (ср. *произвести*), *вытараща*, *воспользуясь* (ср. *вытаращив*, *воспользовавшись*), *по сту* (ср. *по сто*).

Словообразовательные варианты представляют собой варианты образования основ, отклоняющиеся от современных норм: *разоренье* (ср. *разорение*), *авангардия* (ср. *авангард*), *супротивление* (ср. *сопротивление*), *канцелярный* (ср. *канцелярский*), *самодержавство* (ср. *самодержавие*), *напротиву* (ср. *напротив*), *коллегиум* (ср. *коллегия*); *егеров* (ср. *егерей*; подобная форма предполагает твёрдую основу—*егер*).

Фонетические варианты отражают устаревшее произношение слов, в основном заимствованных: *гистория, эсквадра, гранодеры, провинциял-фискал, анбары.*

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*. Это касается и текстов XVIII–XIX вв., и устной речи, и электронной коммуникации, и диалектных текстов. Решение этой проблемы следует искать, по крайней мере, в трех направлениях: 1) нормализация орфографии, 2) пополнение словаря корпуса, 3) обучение программ-парсеров на специфическом для каждого корпуса текстовом материале.

Различия между категориями текстов со значительными отклонениями от литературной нормы состоят в разной степени вариативности и разном соотношении типов вариантов. Поэтому для

каждого корпуса должна избираться наиболее оптимальная тактика работы, учитывающая структуру несловарных единиц. В частности, для корпуса XVIII в., характеризующегося высокой степенью орфографической вариативности, необходима (эффективна) орфографическая нормализация на этапе предварительного технического редактирования и структурной разметки текстов. При таком способе каждому ненормативному написанию приписывается нормативная форма: естли{если*}, зделать{сделать*}, доволно{довольно*} и т.д. В процессе морфологической разметки разбирается нормативная форма, а набор грамматических признаков приписывается всему комплексу, так что при лексико-грамматическом поиске в корпусе на запрос по лемме будут выдаваться контексты, содержащие это слово во всех вариантах написания[4]. Этот путь избран для устных текстов и текстов электронной коммуникации, так что, например, в корпусе устных текстов на запрос «*что*» получаем контексты с *чтo*, *шo*, *чё*[5]. Здесь особую техническую сложность представляют собой колебания «слитное/раздельное написание», учитывая пословный характер принятой в Корпусе разметки. В случае с частотными слитными написаниями конкретных лексических единиц (*когдаб, еслиж*, включая падежные формы—*чегож, чемуб*) можно задать определённые правила и пополнить словарь, но это сложно сделать для текстов с «продуктивным» слитным написанием (аналогичная проблема стоит и для текстов современной электронной коммуникации и частной переписки, где встречаются похожие феномены «неграмотного» письма). Сюда относится уже упоминавшееся «Письмо другу…» Радищева[6], для которого характерно большое количество слитных написаний предлогов (при раздельном написании слов вроде *близ лежащий*): *…Кирасирской Новотроицкой Полк и Киевской пехотной заняли места наблиз лежащих улицах. Все было готово, тысящи зрителей назделанных для того возвышениях и толпа народа разсеяннаго повсем близ лежащим мес-*

[4] Особенно актуальна эта технология для рукописных текстов, например, частных писем, орфография которых может быть весьма далека от нормативной.

[5] См. статью Е. А. Гришиной и С. О. Савчук в наст. сборнике.

[6] В автографах Радищева образцов орфографии вроде *назделанных* как будто не отмечено, так что в данном случае орфография, как можно предполагать, привнесена на стадии печати.

там и кровлям ожидали с нетерпением зрети образ того, котораго предки их в живых ненавидели, а посмерти оплакивали.

Пополнение словаря корпуса предполагает анализ несловарных словоформ и приписывание им грамматических признаков. Для ряда наиболее частотных словоформ, встречающихся и в текстах XIX в., это уже сделано, и они опознаются и размечаются парсером как стандартные формы:

```
<span title="токмо = adv,norm|norm,part">токмо</span>
<span title="кой = acc,apro,inan,norm,pl|apro,nom,norm,pl = r:rel,r:rel">кои</span>
<span title="оный = apro,dat,f,norm,sg|apro,f,gen,norm,sg|apro,f,ins,norm,sg|apro,f,loc,norm,sg = r:dem,r:dem,r:dem,r:dem">оной</span> <span
<span title="нынешний = a,acc,anim,m,norm,plen,sg|a,gen,m,norm,plen,sg = der:adv,r:rel,t:time,der:adv,r:rel,t:time">нынешняго</span>
<span title="прочий = a,acc,inan,norm,pl,plen|a,nom,norm,pl,plen = r:rel,r:rel">прочия</span>
```

Наибольшую сложность представляют морфологические формы, оставшиеся в наследство от старой морфологической системы — «морфологические архаизмы». В современных текстах их можно встретить только в виде застывших осколков в составе фразеологических оборотов (на *босу* ногу, на *круги своя,* темна вода во *облацех*), в то время как в текстах XIX в. их круг достаточно широк [Дич 2005: 93]. В текстах XVIII в., особенно относящихся к первой трети века, старые формы имеют еще более широкое распространение: *пришед* (краткое причастие от *прийти*), формы инфинитива на *–ти* (*восприяти*, *зрети* и под.), *городы*, *домы* (им.-вин. мн.), *детем*, *людем*, *крестьяном* (дат. мн.)[7]. Все эти случаи должны быть включены в состав словаря с соответствующими грамматическими характеристиками.

В дальнейшем варианты — орфографические, морфологические, словообразовательные — могут быть объединены в словаре с соответствующими стандартными формами и образовать словарную

[7] Сложность состоит в том, что некоторые старые формы могут совпадать с современными, и тогда программа-парсер не опознает их как несловарные, а предлагает разборы исходя из нормативной грамматики. Ср.: Сей князь собою видом, как *монстра* span title=”монстр = acc,anim,m,norm,s,sg|anim,gen,m,norm,s,sg = t:hum,r:concr,ev:neg,t:hum,r:concr,ev:neg»> *монстра*

единицу более высокого уровня — гиперлемму. Однако эта гипотеза требует дальнейшей проверки на материале корпуса. Проверка покажет, насколько такое пополнение словаря позволит уменьшить количество ошибочных разборов.

Другой способ снижения шума, который в настоящее время опробуется программистами,—это обучение программы-парсера на подкорпусах однородных текстов (например, разговорных, XVIII-XIX вв.) и настройка таких программ на морфологическую разметку текстов определенного типа. По мнению специалистов, такая настройка позволит программе приписывать словоформе наиболее вероятные разборы.

В заключение остановимся на задачах, которые ставят перед собой разработчики корпуса текстов XVIII в. на ближайшее будущее. Во-первых, это пополнение корпуса новыми текстами, подготовка и включение в состав корпуса редких текстов (частных писем, деловой переписки и записей, старопечатных книг), прошедших процесс соответствующей орфографической обработки. Во-вторых, полный анализ несловарных форм, выделенных в текстах XVIII в. (всего около 3000 словоформ), ручная лемматизация и пополнение словаря корпуса.

Задачей на отдаленную перспективу можно считать создание комплексного информационного ресурса, объединяющего электронную библиотеку оригиналов текстов, представленных в графических форматах, корпус текстов в старой орфографии, которые создаются в Казанском университете [Соловьев, Ахтямов 2006], и корпус текстов в современной орфографии с иными поисковыми возможностями. Такой ресурс мог бы удовлетворить интересы специалистов разных профилей, изучающих культурное наследие XVIII века.

Библиотека — Библиотека электронных ресурсов Исторического факультета МГУ им. М. В. Ломоносова [Электронный ресурс] *http://www.hist.msu.ru/ER/index.html*

Виноградов В. В. Основные этапы истории русского языка // Виноградов В. В. Избранные труды. История русского литературного языка. — М., 1978. — С. 10–64.

Винокур Г. О. История русского литературного языка: Русский литературный язык в первой половине XVIII в. // Избранные работы по русскому языку. — М., 1959. С. 111–137.

Военная — Военная литература [Электронный ресурс] *http://militera.lib.ru*

Восточная — Восточная литература [Электронный ресурс] *http://www.vostlit.info/haupt-Dateien/index-Dateien/H.phtml*

Горшков Н. И. История русского литературного языка. — М., 1969.

Дич Н. Л. О текстах XIX века в национальном корпусе русского языка // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: Индрик, 2005. С. 89–93.

Живов В. М. Язык и культура России XVIII века. М.: Школа «Языки русской культуры», 1996.

Живов В. М. Очерки исторической морфологии русского языка XVII-XVIII веков. — М.: ЯСК, 2004.

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

Национальный корпус русского языка [Электронный ресурс]. — *http://www.ruscorpora.ru*

РВБ — Российская виртуальная библиотека [Электронный ресурс] *http://www.rvb.ru*

Савчук С. О., Сичинава Д. В., Гарипов И. И. Подкорпус текстов XVIII века в составе Национального корпуса русского языка: из опыта работы. *http://fccl.ksu.ru/issue_spec/docs/Savchuk_Sichinava_Garipov.doc*

Савчук С. О., Гришина Е. А. Вариантность в русском языке. Проект словаря // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конфе-

ренции «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14). — М.: РГГУ, 2008. С. 466–474.

Соловьев В. Д., Ахтямов Р. Б. Корпус русского языка XVIII века: текущее состояние // Материалы международной научной конференции. Ижевск, 13–17 июля 2006 г. Ижевск, 2006. С. 156–160.

Успенский Б. А. Из истории русского литературного языка XVIII – начала XIX века. — М., 1985.

ФЭБ — Фундаментальная электронная библиотека «Русская литература и фольклор» [Электронный ресурс] *http://www.feb-web.ru*

Хронос — ХРОНОС [Электронный ресурс] *http://hronos.km.ru*

ImWerden — ImWerden. http:// *www.imwerden.de*.

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