ГРАММАТИЧЕСКИЙ СЛОВАРЬ ДЛЯ АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ XVIII–XIX ВЕКА: ПЕРВЫЕ РЕЗУЛЬТАТЫ[1]

Поляков А. Е. (pollex@mail.ru)
НПБ им. К. Д. Ушинского РАО, Москва, Россия

Савчук С. О. (savsvetlana@mail.ru),
Сичинава Д. В. (mitrius@gmail.com)
Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В статье излагаются основные принципы построения грамматического словаря и морфологического анализатора для текстов XVIII–XIX вв. веков с учётом орфографических, морфологических и лексических особенностей языка этого периода, выявленных на материале текстов Национального корпуса русского языка. Поскольку от данного анализатора требуется универсальность подхода и возможность работы с текстами разных типов и разными орфографическими режимами, то он должен состоять из нескольких модулей, применяемых к текстам различных типов в зависимости от степени проявления в них тех или иных орфографических и грамматических явлений. Его словарь построен на базе существующего грамматического словаря современного русского языка, а также словарей XIX в. и текстов Национального корпуса. Обсуждается несколько альтернативных возможностей реализации орфографических (предобработка, применение технологии параллельного корпуса, нормализация при разметке) и морфологических правил (нормализация при разметке, добавление нестандартных форм в парадигму). Проводится оценка первых результатов применения анализатора к текстам НКРЯ и предлагаются различные варианты улучшения результатов (введение новых правил, пополнение словаря и т. д.).

Ключевые слова: грамматический словарь, автоматический анализ текстов XVIII–XIX вв.

[1] Работа выполнена при поддержке Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика».

A GRAMMAR DICTIONARY FOR AUTOMATIC ANALYSIS OF THE XVIII–XIXTH CENTURY TEXTS: FIRST RESULTS

Polyakov. A. E. (pollex@mail.ru)
Ushinsky's State Scientific Pedagogical Library RAE

Savchuk S. O. (savsvetlana@mail.ru),
Sitchinava D. V. (mitrius@gmail.com)
V. V. Vinogradov's Institute for the Russian Language RAS, Moscow, Russia

The paper presents the key principles of building a grammar dictionary and a morphological analyzer for XVIII–XIXth century Russian texts based on orthographical, morphological and lexical features exemplified by the Russian National Corpus (RNC). The analyzer should involve different modules applicable to different kinds of texts depending on their respective orthographical and grammatical phenomena. Several alternative ways of implementing orthographical and morphological rules are discussed (including pre-processing, online normalization etc.). Evaluation data of the first analysis results are presented.

Key words: grammar dictionary, automatic text analysis, texts of the XVIII–XIXth century

1. Постановка проблемы и способы ее решения

Последнее десятилетие характеризуется ростом интереса к сохранению письменного наследия — текстов предшествующих эпох, в частности, к оцифровке этих текстов и увеличению возможностей доступа и поиска (играет здесь роль и то, что старые тексты, созданные, как минимум, век назад, не охраняются авторским правом, находясь в общественном достоянии). Наблюдается рост числа исторических корпусов, электронных исторических библиотек; объёмы оцифрованных старых изданий, доступных в электронном виде, существенно выросли. Однако основная проблема создателей исторических корпусов и электронных библиотек с поиском по тексту по-прежнему остается открытой. Для морфологического анализа текстов предшествующих эпох (если такая задача вообще ставится) используется анализатор, рассчитанный на современный язык и частично (или даже полностью) на современную орфографию, потому что другого пока нет. Более того, от качества работы такого анализатора зависит уже качество распознавания отсканированного печатного текста в орфографии

XVIII–XIX вв.; словоформы, не предсказываемые современным грамматическим словарём, могут быть распознаны неправильно (например, слово *пылъ* с конечным ером — как *пыль,* а *красныя* — как *красная* или *красный*).

Морфологическая разметка в Национальном корпусе русского языка производится (полу)автоматически с помощью анализаторов Dialing (подкорпус со снятой омонимией) и Mystem (тексты с неснятой омонимией в составе разных подкорпусов). В основе обоих анализаторов лежит грамматический словарь современного русского языка [Зализняк 1977/2003]; словарь анализатора Mystem, по сравнению со словарём Зализняка, пополнен на материале часто встречающихся в Интернете и в поисковых запросах неологизмов 1990–2000-х годов и имён собственных. Результат такой разметки дает значительное количество ошибочных и/или гипотетических разборов. Как было показано в [Савчук, Сичинава 2009], процент погрешностей в разборе текстов XVIII в. находятся на уровне современных текстов, не нормированных (тексты электронной коммуникации) или в принципе не ориентированных на литературную норму (записей диалектной речи).

Повышение качества анализа текстов с большим количеством отклонений от стандарта предлагалось проводить двумя путями: 1) использование стандартного анализатора, отражающего грамматические и орфографические нормы современного литературного языка, с подключением дополнительных правил для отдельных словоформ и категорий слов и 2) использование нового морфологического анализатора, настроенного на определенные тексты.

Между тем база текстов НКРЯ существенно растёт. Расширился объём текстов XVIII в. (к февралю 2012 г. 3,8 млн слов в прозаическом подкорпусе плюс почти 1 млн в поэтических текстах) и памятников предшествующего периода — среднерусских текстов XIV–XVII веков (3 млн слов). Увеличение объема исторических текстов, а также их хронологического разнообразия делает более актуальным использование нового анализатора. Морфологической анализатор церковнославянского языка, разрабатываемый для церковнославянского подкорпуса НКРЯ [Поляков и др., 2012], для этой цели в текущем виде не подходит (его пробное применение к среднерусским текстам, без орфографической настройки, показало, что опознаётся лишь примерно половина словоформ).

2. Специфика корпуса исторических текстов

Для среднерусского периода было характерно сосуществование текстов с более сильными церковнославянскими либо народными тенденциями; представление о «диглоссии» — жёстком распределении жанров между двумя языками — несколько упрощено, хотя и отражает определённые установки писавших. Явлением более или менее тесного взаимопроникновения двух систем, их «гибридизации», собственно говоря, и объясняется необходимость учета элементов лексики и грамматики фактически другого языка при анализе русских текстов.

Такое сосуществование, хотя и в меньшей степени, сохранялось и позже, в XVIII—начале XIX вв. Тексты разных периодов могут значительно отличаться по характеру языка, поскольку конец XVII—середина XIX в — это самый интенсивный период формирования литературного языка нового типа; различные авторы привлекают различные разговорные формы, заимствования, по-разному передают имена собственные и т. п. Важной чертой данного периода (по крайней мере до поколения Пушкина и Карамзина) является существенная орфографическая пестрота.

До недавнего времени Национальный корпус русского языка включал ранние тексты (XVIII — первая половина XX в.; о среднерусском периоде речь не идёт) только в современной орфографии, или, по крайней мере, с некоторыми отклонениями от современной нормы, но в целом ориентированные на орфографический режим 1918 г. или даже 1956 г. [Савчук, Сичинава, Гарипов 2006]. Это диктовалось, помимо традиции, и ориентацией на авторитетные научные издания советского и постсоветского времени. Вместе с тем давно обсуждается вопрос о включении в корпус значительного количества текстов в дореформенной орфографии [Соловьев, Ахтямов 2006; Савчук 2008], с сохранением упразднённых реформой 1918 года графем и орфографических правил, тем более что анализатор Mystem умеет учитывать большинство этих правил (использование *ѣ, ѳ, і*, конечного *ъ*, окончания *-ыя, -ія, -аго, -яго*). Вместе с тем встаёт проблема обработки и совместного поиска по текстам разных типов. Например, для основного корпуса НКРЯ желательна возможность индексировать корпус так, чтобы при поиске точных форм можно было найти одновременно и дореформенное, и современное написание (например, чтобы по точному запросу словоформы *пѣной* находилась бы и словоформа *пеной* в новоорфографических текстах).

3. Принцип работы анализатора. Составные элементы анализатора

Можно сформулировать требования, которым должен отвечать анализатор для обработки исторических текстов:

А) *Универсальность*: анализатор должен уметь работать с текстами, обладающими различными характеристиками:

1) обрабатывать тексты в новой и дореформенной орфографиях, учитывая информацию, которая несёт каждая из них (например, не просто приравнивать омофоничные буквы, но и отличать *всѣ* от *всё*);
2) обрабатывать тексты разных жанров от религиозных, написанных на церковнославянском или приближенном к нему, до бытовых писем в свободной орфографии (далекой от нормализации).

Б) *Открытость* — способность пополняться и видоизменяться, настраиваться на разные типы текстов, возможность «обучаться» на основании пополненного словаря и т. п.[2]

Несмотря на требование универсальности, одновременно могут сосуществовать и модификации анализатора для текстов разных периодов и/или жанров, использующие ряд правил, специфических именно для этих текстов. Например, к текстам XX–XXI вв. едва ли нужно применять правило, согласно которому частицы *б(ы), ж(е)* и *ли/ль* могут писаться слитно с предыдущим словом (что особо часто встречается в XVIII в.); это приведёт исключительно к паразитическим разборам типа *мыль = мы + ль, стали = ста + ли, ниже = ни + же* и др. (в текстах XVIII в. разборы такого рода можно отсеивать при помощи специального правила). Аналогично, такие специфические для письменности раннего XVIII в., а также XVII и предшествующих веков правила, как пропуск мягкого знака *(толко = только)* и отсутствие в графической системе буквы *й* (*таино = тайно*) приведёт к излишним неправдоподобным разборам в современных текстах (типа *банка = банька, заика = зайка* и т. п.). Список периодов и жанров, требующих особых модификаций анализатора, нужно будет установить опытным путём после анализа текстов, входящих в Национальный корпус русского языка. В частности, можно предположить, что различные модификации потребуются для текстов бытовых грамоток XVII в., для текстов разных периодов, ориентированных на церковнославянский язык, для собственно русских текстов XVIII в., первой половины XIX в. и нескольких дальнейших периодов.

Исходя из этих требований, перед разработчиками стоят два типа задач.

1. *Грамматические*: разработка грамматических парадигм для лексем, отсутствующих в современном словаре.

2. *Орфографические*: обеспечение лемматизации форм, имеющих отклонения от стандартных написаний. К решению второй (орфографической) задачи имеется ряд возможных подходов.

1) Предобработка (preprocessing) — нормализация теста, своего рода «перевод» его на стандартный язык, предваряющая морфологический анализ. Нормализация текста широко применяется в устных и диалектных корпусах разных языков и в подкорпусе электронной коммуникации НКРЯ [см. об этом Гришина, Савчук 2009] в ручном режиме. Ср. конкретные примеры применения такой технологии: [Качинская 2010] (диалектный корпус), [Baron, Raison 2009], [Lay 2012] (исторические корпуса). Недостатки метода предобработки заключаются в значительной затрате ресурсов для подготовки нормализованной версии текста и зачастую неоднозначности такой нормализации при объёме корпуса в несколько миллионов слов.

---

[2] В отношении несовременных текстов (представляющих собой по определению закрытый, хотя и очень большой класс) о принципе открытости анализатора можно говорить с известной долей условности, однако это обстоятельство не играет практической роли до тех пор, пока все или почти все исторические тексты не будут включены в корпус. В настоящее время мы еще очень далеки от этого (достаточно сказать, что не только оцифровано, но и вообще издано лишь незначительное меньшинство сохранившихся от XVII–XVIII вв. текстов, в том числе и бытовых текстов в нестандартной орфографии).

2) Применение технологий параллельного корпуса — одновременное использование как оригинального текста, так и перевода орфографии на современные нормы, выровненных по предложениям или даже словоформам [Meyer 2009]. Национальный корпус русского языка уже поддерживает параллельную технологию — выравнивание текстов по предложениям, которая используется, прежде всего, в двуязычных и многоязычном параллельных подкорпусах НКРЯ (ср. [Добровольский, Кретов, Шаров, 2005], [Sitchinava 2012]). Фактически эта технология представляет собой расширение предыдущей: нормализованный текст становится доступен пользователю.
3) Учет различных уровней вариативности в грамматическом словаре анализатора. Именно данный подход лежит в основе предлагаемого в настоящей публикации. В грамматический словарь анализатора, выстроенный на базе лексикографических источников, дополнительно включаются словоформы, не распознаваемые или распознаваемые неправильно и неполно существующими анализаторами. Применение данного подхода к анализу конкретных текстов разных типов и периодов покажет, в какой степени он позволит сократить количество неправильных разборов, и, возможно, обнаружит участки, в принципе не поддающиеся автоматическому анализу и требующие ручной нормализации.

4. Грамматический словарь

4.1. Общие принципы

Общие принципы и алгоритм работы анализатора, области его применения были изложены в [Поляков 2012]. Грамматический словарь определяется как список лексем языка с приписанной информацией о словоизменении. Каждая лексема в словаре содержит, как минимум, следующую информацию: 1) основа с указанием чередований; 2) постоянные признаки лексемы (часть речи, род, одушевленность, переходность, и т.д.); 3) код словоизменительного типа (парадигмы). Вот пример записи для некоторых глаголов с чередованиями в основе:

но(с|ш)+ить	V,ipf,tr	V4
б(и|ь|е)+ть	V,ipf,tr	V11
пе(к|ч|)+ь	V,ipf,tr	V8
ж(г|ж|ег|е)+чь	V,ipf,tr	V8*g

Грамматический словарь анализатора складывается из нескольких модулей, соответствующих различным периодам истории русского языка. При анализе конкретного текста выбирается модуль, соответствующий типу и периоду создания текста.

1) Современный модуль строится на основе Грамматического словаря Зализняка [Зализняк 1977/2003], который фиксирует лексический

и грамматический стандарт конца XX века. Тем не менее, этот модуль позволяет анализировать значительную часть словоформ, встречающихся в текстах XVIII–XIX века, если они не имеют существенных орфографических и грамматических отличий от современной нормы.

2) Модуль XVIII–XIX века строится на основе анализа корпуса реальных текстов, а также исторических словарей русского языка, включая:
   - Словарь Академии Российской (1789–1794);
   - Словарь церковнославянского и русского языка (ЦСРЯ) (1847);
   - Полный русский орфографический словарь (1898);
   - Словарь русского языка XVIII века.
3) Модуль XVII века и более ранних периодов строится в основном на основе анализа корпуса реальных текстов и лишь частично на основе исторического словаря (Словарь русского языка XIV–XVII века).

Для адекватного анализа текстов XVIII–XIX века, часть которых представлена в дореформенной орфографии, необходимо добавить в грамматическую модель анализатора следующие формы:

1) формы, характерные для всех периодов:
   - деепричастия совершенного вида от основы презенса (*прийдя, увидя, взгромоздясь*), которые вполне употребительны в современном языке, а тем более в языке XVIII–XIX века;
   - нестандартное распределение *-ся/-сь* после гласных (*валюся, валилася; запершегось),* существующее как в современном языке (в некоторых идиолектах), а также в языке XVIII–XIX века;
   - сравнительная степень на *-ей* (*сильней*) и с префиксом *по-* (*посильнее, посильней*);
2) формы, характерные для XIX века и более ранних периодов:
   - адъективные флексии (*-аго/-яго, -ыя/-ія*);
   - особые формы местоимений (*ея, онъ, однъ, однъхъ*);
   - творительный падеж 3-го склонения на *-ію* (*милостію, помощію*);
3) формы, характерные для XVIII века и более ранних периодов:
   - сравнительная степень на *-яе* (*сильняе, скоряе*);
   - глагольные флексии *-ти* и *-ши* (*ходиши, ходити*), которые, скорее всего, должны трактоваться как церковнославянизмы (см. ниже).
4) церковнославянские формы:
   - формы имперфекта (*творяше, творяхомъ, творяху*) и аориста (*творихъ, творихомъ, твориша*), которые нередко бывают омонимичны (*делахъ, делахомъ*);
   - частотные формы косвенных падежей существительных (*градомъ, градѣхъ, отроцы, отроцѣхъ*);
   - частотные лексемы (*иже*, *яко, понеже, вельми*);

и т. д.

Анализатор, помимо словаря, сопровождается отдельными правилами для анализа текстов в разных орфографических режимах — алгоритмической нормализацией, одновременной с приписываемой морфологической разметкой. Они работают по следующему принципу: если словоформа не предсказывается на основании грамматического словаря (а может получить только гипотетический разбор), то предпринимаются попытки, исходя из её буквенного состава, регулярной замены в ней тех или иных букв и анализа получившейся нормализованной словоформы. Если эта словоформа получает негипотетический анализ (и, возможно, этот анализ удовлетворяет тем или иным грамматическим ограничениям), то он подставляется в качестве её разбора. Таковы графические (типа *ѳ => ф,* см. ниже, 4.3.А) и орфографические (типа *цы => ци*, см. ниже, 4.3.Б) правила, применимые ко всем словоформам, удовлетворяющим данным критериям.

Вместе с тем должен быть задействован также ряд индивидуальных правил, вводящих конкретные орфографические варианты для конкретных лемм; например, такова индивидуальная вариативность типа *естьли~если* (неизменяемое), *потчевать~подчивать, ветчина~вядчина* (проводится по всей парадигме).

4.2. Источники для формирования словника грамматического словаря

Источниками для формирования словника грамматического словаря являются, с одной стороны, существующие исторические словари русского языка (см. выше, п. 4.1), с другой — результаты анализа корпуса реальных текстов. Лексемы из этих двух источников будут добавляться к основному модулю, составленному на основе Грамматического словаря Зализняка. Этот путь представляется нам самым коротким для достижения практических целей — адекватного анализа текстов исторического корпуса.

Корпусной словник создается на базе частотного списка словоформ, извлеченных из текстов, которым при автоматическом анализе приписываются леммы и грамматические характеристики [Савчук 2012]. Пополнение словника грамматического словаря из обоих источников будет осуществляться поэтапно, по мере подготовки электронных версий словарей, одной стороны, и развития корпуса исторических текстов, с другой.

4.3. Методы анализа вариативности

Для правильного анализа и сведения к единой лемме языковых вариантов на различных уровнях анализатор может использовать правила следующего типа.

А) *Графические правила*

Графические правила основаны на приравнивании графем, встретившихся в тексте, графемам, входящим в порождаемые словарём формы, например, *ѳ => ф; ѯ => кс*.

Б) *Орфографические правила*

Орфографические правила связаны с нормализацией определённых орфограмм: иными словами, определённые буквы заменяются на другие лишь в контексте некоторых третьих, при этом обе буквы присутствуют в алфавите грамматического словаря. Таково, например, правило *цы => ци*, при соблюдении которого стандартизированный разбор получают написания типа *цыновка, цыдулка, цыгарка, цыгейка*, нормативные до 1956 г.

Пример работы орфографического правила:

в тексте представлена словоформа *надеютца;*
из-за наличия конечного *тца* проверяется правило *тца=ться/тся*;
словоформа *надеються* словарем не порождается;
словоформа *надеются* словарем порождается и разбирается как lex=НАДЕЯТЬСЯ gr=praes,3p,pl;
словоформа *надеютца* получает разбор lex=НАДЕЯТЬСЯ gr=praes,3p,pl=distort.

Аналогично словоформа *носитца*, для которой возможны две нормализации, получает два разбора — инфинитивный *(носиться)* и финитный *(носится),* а словоформа *отца* только разборы от слова ОТЕЦ (ввиду отсутствия словоформ **от(ь)ся*).

В) *Морфологические правила*

Морфологические правила, в отличие от орфографических, устроены с учётом информации о грамматическом разборе словарной словоформы. Например, вводится правило, которое можно записать как *-яе*, comp,anom <= *-ее,* comp. Это значит, что если в тексте встретилась не получающая словарного анализа словоформа, кончающаяся на *-яе*, например, *сильняе*, а замена *-яе* на *-ее* в этой словоформе даёт словарную форму (*сильнее*), и притом эта форма есть словоформа сравнительной степени, то словоформа типа *сильняе* получает такой же грамматический разбор, что и словоформа типа *сильнее* плюс помету anom («аномальная морфологическая форма»).

Г) *Списочный способ* — задание вариативности списками, на ограниченных классах единиц. Таково, например, орфографическое правило, согласно которому ряд конкретных корней может (или даже практически обязан) в среднерусских и церковнославянских текстах сокращаться (записываться под титлом), ср. такие словоформы, как *Г(о)с(по)дь, м(е)с(я)ц, гл(аголе)ть* и т. д. К списочным правилам относятся и блокирующие правила для частотных паразитических разборов типа *стали = ста ли*.

Указанные правила имет определенный порядок применения; так, в первую очередь применяются графические правила, в том числе списочные; затем — орфографические, морфологические и списочные правила, блокирующие паразитические разборы.

В процессе формирования словаря и оптимизации работы анализатора предусматривается несколько циклов обработки текстов. Каждая новая версия анализатора будет проходить проверку на корпусе: после анализа результатов в словарь и правила анализатора будут вноситься пополнения и коррективы, после чего этот цикл повторяется уже с новой версией. Далее мы изложим оценку результатов первой версии анализатора на корпусе текстов XVIII–XIX веков.

5. Оценка результатов

5.1. Состав несловарных словоформ

Первый вариант анализатора на основе первой версии словаря был опробован на экспериментальном корпусе объемом около 4 млн словоупотреблений[3], который включает 256 тыс. различных словоформ. Результаты представлены в таблице.

разобрано	185 221	72,4 %
гипотезы	63 904	25,0 %
не разобрано	6 780	2,6 %

Наибольший интерес для дальнейшей разработки словаря представляет анализ словоформ, которые не были опознаны как слова русского языка или получили гипотетические разборы, и оценка предложенных гипотез. Список неопознанных форм в настоящее время полностью проанализирован, всем формам вручную приписаны леммы, и они пополнили список разобранных форм. Перечислим наиболее массовые случаи.

Не распознаны или неправильно опознаны сочетания знаменательных частей речи с частицами *-то(-та, -ат), же(ж), ли(ль), бы(б), -де, -ка*, в написании которых в разные периоды наблюдались колебания. В изданиях XVIII в. в написаниях частиц не было последовательности, можно встретить дефисные, слитные и раздельные написания: *них-же, месте-же, нихже, мыже, таковаже, пили-б, ожидали-б, где-б, пилиб, ожидалиб, еслиб.* В НКРЯ сочетания с раздельным и дефисным написанием частиц анализируются как две леммы, при этом автоматический анализ в большинстве случаев правильный. Эти же решения следует использовать и в исторической части словаря.

---

[3] Экспериментальный корпус составлен на основе текстов XVIII — первой трети XIX вв., входящих в состав НКРЯ. По жанровому составу корпус XVIII в. разнообразен: доля художественных текстов и публицистики — по 24 %, церковно-богословские тексты составляют 19 %, научные тексты ф — 17 %, официальные документы — 11 %, бытовые тексты (письма, дневники) — 5 %. Приблизительно в тех же пропорциях представлены и тексты XIX в. Хронологически тексты экспериментального корпуса распределяются следующим образом: 1700–1730 — 6 %, 1731-1780 — 43 %, 1781–1799 — 30 %, 1800–1830 — 21 %. Подробнее о составе корпуса XVIII в. см. [Савчук, Сичинава 2009]).

Другую большую группу словоформ, не получивших разборов, составляют архаические формы склонения и спряжения. Среди них заметное место занимают: существительные, прилагательные в форме тв. п. на *-ою* (263 формы): *Гришкою, Кабардою, прежестокою*; причастия в форме им. п. мн. ч. на *-ии* (101) / *-ыи* (18): *входящии*, *нарицающии*, *приежжаемыи*, *украшенныи*; причастия на *-яй* (128): *возвышаяй*, *вступаяй*; краткие причастия на *-ущ / ащ* (27): *блистающ*, *властвующ*; формы имперфекта: *живяше*, *знаяше* (10); *бяху, стояху, мняху* (22).

Третью многочисленную группу составляют орфографические варианты: *безщчетну, возмеш, баталиах, полицыи, прокломации, поосчрении* (написания с *сч* вместо *щ* совершенно регулярны, например, у Татищева, который в своем орфографическом трактате утверждал об избыточности буквы *щ* и свою фамилию писал как *Татисчев*) и др. Здесь анализ отдельных групп вариантов приведет к формулированию частных формальных правил анализа соответствующих орфограмм (см. п. 4.3).

Четвертую группу составляют многочисленные собственные имена — топонимы, имена, фамилии и отчества лиц, литературных героев и мифологических персонажей, причем многие из них присутствуют в текстах в нескольких вариантах: *Шлиссельбург, Шлюссембурх, Шлютенбурх, Шлютелбург, Шлютельбург, Слюсинбург, Слютелбург, Слютельбург, Валпарейсо, Валпарейзо, Вальпарейзо* (в современной передаче — *Вальпараисо*, город в Чили), *Елизавета, Елисавета, Елисавет, Елисавефь, Елисафет, Ньютон, Нейтон, Невтон, Дон-Кихот, Дон-Кишот, Донкишот, Микель-Анджело, Мишель Анжело* и др. Примыкают к этой группе производные от собственных имен — прилагательные и существительные: *европскии, коперниканскии, ефесския* (здесь играют роль и архаичные окончания), *Антошка, Бомонтша* и т. д.

5.2. Оценка гипотетических разборов

Среди словоформ, получивших гипотетические разборы, можно выделить зоны с высоким уровнем предсказуемости (25–30%, то есть одна гипотетическая лемма из трех или четырех предложенных оказывается правильной), зоны средней предсказуемости (предлагается от пяти гипотез, из них одна правильная) и зоны с нулевой предсказуемостью (ни одна из предложенных гипотез не является правильной).

Примеры высокой предсказуемости обнаружили, в частности, существительные с основой на *-к-* (около 4% словоформ, получивших варианты разборов), спрягаемые формы глаголов, за исключением архаичных (более 6% словоформ), формы прилагательных (около 35% всех словоформ с гипотетическими разборами):

доимка доимка?=N,f,inan=sg,nom=N33* | доимок?=N,m,inan=sg,gen=N13* | доимка?=N,f,anim=sg,nom=N33*

напояют напоять?=V,ipf=ind,pres,pl,3,act=V1 | напоять?=V,ipf,intr=ind,pres,pl,3,act=V1 | напоять?=V,pf=ind,fut,pl,3,act=V1

Из архаических форм в зоне высокой предсказуемости находятся формы прилагательных мн. ч им. п. на *-ия /-ыя, -аго*:

```
одинакия одинакий?=A=pl,nom/acc=A3 | одинакий?=A=pl,nom/acc=A3*
    | одинакая?=N,f,inan=pl,nom/acc/acc=A3 | одинакой?=A=pl,nom/
    acc=A3b | одинакое?=N,n,inan=pl,nom/acc/acc=A3
```

Пример средней степени предсказуемости:

```
неведь неведь?=N,f,inan=sg,nom/acc=N41 |
    неведь?=N,m,anim=sg,nom=N12 | неведь?=N,topn,m,anim/
    inan=sg,nom=N12 | неведь?=CONJ/PART | неведь?=PREP
```

Наибольшие сложности представляет идентификация архаических глагольных форм. Все они попадают в зону нулевой предсказуемости. Причина объяснена выше — отсутствие в словнике в достаточном объеме церковнославянских глаголов и парадигм спряжения.

Список словоформ с гипотетическими разборами находится в стадии обработки: словоформам с ошибочными разборами вручную приписываются леммы и грамматические признаки, для форм со средней предсказуемостью рассматриваются способы сокращения предлагаемых гипотез.

5.3. Сокращение количества гипотез

Сокращения предлагаемых гипотез можно добиться путем использования правил, которые будут применяться к определенным классам словоформ, образующим закрытые или пополняемые списки. Часть правил будет основана на учете связи морфологических признаков с морфемной структурой слова. Так, например, для форм на *-ения* анализатор предлагает 5 гипотетических лемм:

```
посмотрения посмотрение?=N,n,inan=pl,nom/acc¦sg,gen=N27 | посмот
    рения?=N,f,inan=sg,nom=N37 | посмотрения?=N,topn,f,inan=sg,no
    m=N37 | посмотрения?=N,persn,f,anim=sg,nom=N37 | посмотрений?
    =N,m,inan=sg,gen=N17
```

При этом статистически формы распределяются следующим образом. Из 444 форм 439 (98%) являются формами род. п. существительных среднего рода, 2 формы — род. п. мужского рода и 3 — им. п. женского рода (имена собственные). Следовательно, формам на *-ения* целесообразно приписывать леммы существительных среднего рода как наиболее статистически значимые, а существительные женского и мужского рода задать списком и включить в словарь.

Другой источник сокращения количества гипотез, как уже говорилось в п. 4.3, видится в анализе орфографической вариативности, реально

представленной в текстах, и применении орфографических преобразований. Приведем несколько примеров.

Слитное написание *не* с глаголами: *небудет, неисповедует, ненайдет, нетребует*;

сч => щ: *поосчрение=поощрение, немосчный=немощный* и др., более 500 форм;

жы, шы => жи, ши: *показавшые, вашы, стражы, лжы* (55);

Список орфографических особенностей текстов в старой орфографии см. также в [Поляков 2012].

6. Заключение

Принципы формирования грамматического словаря для автоматического анализа текстов XVIII–XIX вв., описанные в настоящей статье, были опробованы на экспериментальном корпусе, составленном из текстов XVIII — 1-ой трети XIX в. Результаты автоматической морфологической разметки показали, что более 73 % словоформ корпуса получили однозначные разборы, но такой сравнительно высокий показатель связан с тем, что в корпусе пока мало текстов начала XVIII в., а кроме того, большая часть текстов, в соответствии с ориентацией на авторитетные научные издания советского и постсоветского времени, представлена в современной орфографии или с некоторыми отклонениями от современной нормы. Очевидно, что увеличение доли оригинальных текстов XVII и первой трети XVIII в. изменило бы результат анализа в сторону уменьшения доли совпадающих словоформ и привело бы к росту количества ошибочных разборов.

Изучение и классификация ошибок анализатора позволило наметить пути дальнейшего расширения и совершенствования словаря. В ближайшие планы входит внедрение всех изменений в словарь, тестирование его новой версии на экспериментальном корпусе, откорректированном с учетом выявленных ошибок, а также на отдельных текстах более раннего периода, относящихся к разным жанрам.

Литература

1. *Гришина Е. А., Савчук С. О.* Корпус устных текстов в НКРЯ: состав и структура // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 129–149.
2. *Добровольский Д. О., Кретов А. А., Шаров С. А.* Корпус параллельных текстов: архитектура и возможности использования. // Национальный корпус русского языка: 2003–2005. — М.: Индрик, 2005. С. 263–296.
3. *Зализняк* 1977/2003 — Зализняк А. А. Грамматический словарь русского языка. Изд. 1-е. М., 1977 (4-е изд., испр. и доп., М. 2003)
4. *Качинская И. Б.* Диалектный подкорпус НКРЯ. Новый стандарт подачи. Новое рабочее место // О. Ю. Крючкова и др. (ред.) Русская устная речь. Материалы

международной научной конференции «Баранниковские чтения. Устная речь: русская диалектная и разговорно-просторечная культура общения» и межвузовского совещания «Проблемы создания и использования диалектных корпусов». Саратов, Издательский центр «Наука», 2011. С. 245–255

5. *Поляков А. Е.* Проблемы и методы анализа русских текстов в дореформенной орфографии // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (2012). Вып. 11. М., 2012. С. 536–547.
6. *Поляков А. Е., Добрушина Е. Р., Иванова-Алленова Т. Ю.* Корпус церковнославянских текстов в составе НКРЯ, первая версия: проблемы и решения. Информационные технологии и письменное наследие. Материалы международной научной конференции. — Петрозаводск, 2012. С. 211–215.
7. *Савчук С. О., Сичинава Д. В., Гарипов И.* Подкорпус текстов XVIII века в составе Национального корпуса русского языка: из опыта работы // Web Journal of Formal, Computational & Cognitive Linguistics. Специальный выпуск (Труды Российского научно-образовательного центра по лингвистике им И. А. Бодуэна де Куртенэ), 2006.
8. *Савчук С. О.* Корпус русских текстов XVIII века в составе Национального корпуса русского языка: проблемы и перспективы // Информационные технологии и письменное наследие. Материалы международной научной конференции. Казань, 2008. С. 241–244.
9. *Савчук С. О., Сичинава Д. В.* Корпус русских текстов XVIII века в составе НКРЯ: проблемы и перспективы // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 52–70.
10. *Савчук С. О.* Электронный словарь вариантов на основе текстов XVIII в. Информационные технологии и письменное наследие. Материалы международной научной конференции. — Петрозаводск, 2012. С. 241–244.
11. *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.
12. *Baron, A., Raison, P.* (2009) Automatic standardization of texts containing spelling variations. How much training data do you need? // In M. Mahlberg, V. González-Díaz, and C. Smith (eds.), Proceedings of the Corpus Linguistics Conference, CL2009. University of Liverpool, UK, pp. 20–23, http://ucrel.lancs.ac.uk/publications/CL2009/314_FullPaper.pdf
13. *Lay, M. H.* (2012) VariaLog: how to locate words in a French Renaissance Virtual Library // Digital Humanities Conference, University of Hamburg, Germany, 2012, available at: http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/varialog-how-to-locate-words-in-a-french-renaissance-virtual-library/
14. *Meyer, R.* (2009) Semi-automatic morphosyntactic tagging of a diachronic corpus of Russian // In M. Mahlberg, V. González-Díaz, and C. Smith (eds.), Proceedings of the Corpus Linguistics Conference, CL2009. University of Liverpool, UK, pp. 20–23. http://ucrel.lancs.ac.uk/publications/cl2009/abstracts.htm#347
15. *Sitchinava D.* (2012) Parallel corpora within the Russian National Copus // *Prace Filologiczne*, LXIII, 2012. С. 271–278.

References

1. *Baron, A., Raison, P.* (2009), Automatic standardization of texts containing spelling variations. How much training data do you need? M. Mahlberg, V. González-Díaz, and C. Smith (eds.), Proceedings of the Corpus Linguistics Conference, CL2009. University of Liverpool, UK, pp. 20–23, available at: http://ucrel.lancs.ac.uk/publications/CL2009/ 314_FullPaper.pdf
2. *Bolshakov, I. A., Bolshakova, E. I.* (2012), An Automatic morphological classifier of noun phrases in Russian [Avtomaticheskij morfoklassifikator russkih imennyh grupp]. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialog" 2012 [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Po materialam ezhegodnoj mezhdunarodnoj konferencii "Dialog" 2012]. Bekasovo, pp. 81–92.
3. *Dobrovol'skij D. O., Kretov A. A., Sharov S. A.* (2005), Parallel Corpus: architecture and usability [Korpus parallel'nyh tekstov: arhitektura i vozmozhnosti ispol'zovanija], in Russian National Corpus: 2003–2005 [Nacional'nyj korpus russkogo jazyka: 2003–2005], Indrik, Moscow, pp. 263–296.
4. *Grishina E. A., Savchuk S. O.* (2009), Spoken texts in the RNC: composition and structure [Korpus ustnyh tekstov v NKRJa: sostav i struktura], in Russian National Corpus: 2006–2008. New Results and Perspectives [Nacional'nyj korpus russkogo jazyka: 2006–2008. Novye rezul'taty i perspektivy]. Nestor-Istorija, SPb, pp. 129–149.
5. *Grot Ja. K.* (1873) Controversial issues of Russian spelling since Peter the Great until now [Spornye voprosy russkogo pravopisanija ot Petra Velikogo donyne]. Tipografija imperatorskoj AN, SPb.
6. *Kachinskaja I. B.* (2011), Dialectal subcorpus of the RNC. The new standards. New workplace [Dialektnyj podkorpus NKRJa. Novyj standart podachi. Novoe rabochee mesto], in O. Ju. Krjuchkova i dr. (red.) Russian speech. Proceedings of the International Conference "Barannikovskie reading. Spoken speech: Russian dialect and colloquial vernacular culture of communication" and intercollegiate conference "Development and Use of dialect corpora" [Russkaja ustnaja rech'. Materialy mezhdunarodnoj nauchnoj konferencii "Barannikovskie chtenija. Ustnaja rech': russkaja dialektnaja i razgovorno-prostorechnaja kul'tura obshhenija" i mezhvuzovskogo soveshhanija "Problemy sozdanija i ispol'zovanija dialektnyh korpusov"], Izdatel'skij centr "Nauka", Saratov, pp. 245–255.
7. *Lay, M. H.* (2012), VariaLog: how to locate words in a French Renaissance Virtual Library, Digital Humanities Conference, University of Hamburg, Germany, 2012, available at: http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/varialog-how-to-locate-words-in-a-french-renaissance-virtual-library/
8. *Meyer, R.* (2009), Semi-automatic morphosyntactic tagging of a diachronic corpus of Russian, M. Mahlberg, V. González-Díaz, and C. Smith (eds.), Proceedings of the Corpus Linguistics Conference, CL2009. University of Liverpool, UK, pp. 20–23, available at: http://ucrel.lancs.ac.uk/publications/cl2009/abstracts.htm#347

9. *Poljakov A. E.* (2012), Problems and methods in analysis of Russian texts in the pre-reform spelling [Problemy i metody analiza russkih tekstov v doreformennoj orfografii], Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialog" 2012 [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Po materialam ezhegodnoj mezhdunarodnoj konferencii "Dialog" 2012]. Bekasovo, pp. 536–547.
10. *Poljakov A. E., Dobrushina E. R., Ivanova-Allenova T. Ju.* (2012), Corpus of Church Slavonic texts in the RNC, the first version: problems and solutions. [Korpus cerkovnoslavjanskih tekstov v sostave nkrja, pervaja versija: problemy i reshenija], Information technology and the written heritage. Proceedings of the International Conference [Informacionnye tehnologii i pis'mennoe nasledie. Materialy mezhdunarodnoj nauchnoj konferencii], Petrozavodsk, pp. 211–215.
11. *Savchuk S. O., Sichinava D. V., Garipov I.* (2006), Subcorpus of the XVIIIth century texts in the Russian National Corpus [Podkorpus tekstov XVIII veka v sostave Nacional'nogo korpusa russkogo jazyka: iz opyta raboty], Web Journal of Formal, Computational & Cognitive Linguistics. Special Issue (Proceedings of the Baudouin de Courtenay Russian Research and Educational Center in linguistics) [Special'nyj vypusk (Trudy Rossijskogo nauchno-obrazovatel'nogo centra po lingvistike im. I. A. Boduena de Kurtene)], available at: http://fccl.ksu.ru/fcclpap.htm.
12. *Savchuk S. O.* (2008), Corpus of the Russian XVIIIth century texts in the Russian National Corpus: problems and prospects [Korpus russkih tekstov XVIII veka v sostave Nacional'nogo korpusa russkogo jazyka: problemy i perspektivy]. Information technologies and written heritage. Proceedings of the International Conference [Informacionnye tehnologii i pis'mennoe nasledie. Materialy mezhdunarodnoj nauchnoj konferencii], Kazan, p. 241–244.
13. *Savchuk S. O., Sichinava D. V.* (2009), Corpus of the Russian XVIIIth century texts in the RNC: Problems and Perspectives [Korpus russkih tekstov XVIII veka v sostave NKRJa: problemy i perspektivy], in Russian National Corpus: 2006–2008. New Results and Perspectives [Nacional'nyj korpus russkogo jazyka: 2006–2008. Novye rezul'taty i perspektivy], Nestor-Istorija, SPb, pp. 52–70.
14. *Savchuk S. O.* (2012), Electronic dictionary of variants based on the 18th century texts [Elektronnyj slovar' variantov na osnove tekstov XVIII v.], Information technology and the written heritage. Proceedings of the International Conference [Informacionnye tehnologii i pis'mennoe nasledie. Materialy mezhdunarodnoj nauchnoj konferencii], Petrozavodsk, pp. 241–244.
15. *Sitchinava D.* (2012), Parallel corpora within the Russian National Corpus. *Prace Filologiczne*, LXIII, 2012, pp. 271–278
16. *Solovyev V. D., Akhtyamov R. B.* (2006), Corpus of the XVIIIth century Russian: the present state of affairs. [Korpus russkogo jazyka XVIII veka: tekushhee sostojanie]. Proceedings of the International Conference [Materialy mezhdunarodnoj nauchnoj konferencii]. Izhevsk, pp. 156–160.
17. *Zaliznjak, A. A.* (1977/2003), Grammatical dictionary of the Russian language [Grammaticheskij slovar' russkogo jazyka], Moscow (4 ed. 2003).