ПРОБЛЕМЫ И МЕТОДЫ АНАЛИЗА РУССКИХ ТЕКСТОВ В ДОРЕФОРМЕННОЙ ОРФОГРАФИИ[1]

Поляков А. Е. (pollex@mail.ru)

ГНПБ им. К. Д. Ушинского, Москва, Россия

Существующие лингвистические процессоры не пригодны для анализа текстов в дореформенной орфографии из-за многочисленных графических, морфологических и лексических отличий языка 18–19 века от современного языка. Мы разработали лемматизатор, который умеет правильно анализировать тексты в дореформенной орфографии, а также включает возможность гибкой настройки на другие орфографические системы (включая смешанную орфографию). В данной работе рассматриваются проблемы, возникающие при анализе русских дореформенных текстов, и возможные пути их решения.

Ключевые слова: словоизменение, русский язык, дореформенная орфография, автоматический анализ

[1] Данное исследование выполнено в рамках работ по гранту РФФИ 11-06-00197 «Создание программного модуля проверки русской дореформенной орфографии».

PROBLEMS AND METHODS IN ANALYSIS OF RUSSIAN TEXTS IN PRE-REFORM SPELLING

Polyakov A. E. (pollex@mail.ru)

Ushinsky State Scientific Pedagogical Library, Moscow, Russia

The existing linguistic processors (spellcheckers, lemmatizers, OCR programs) are not suitable for analysis of pre-reform Russian texts because of numerous graphical, morphological and lexical differences from the modern Russian language. Some lemmatizers have a restricted support of pre-reform spelling, but they are closed source and cannot be modified or extended. We have developed a lemmatizer which can properly analyze pre-reform texts and has a facility for flexible adaptation to other spelling systems.
This paper discusses the problems in the analysis of pre-reform Russian texts (obsolete forms, spelling variants) and the methods of their solution (normalization, modification of the grammatical model, etc.).

Key words: inflection, Russian language, pre-reform spelling, automated analysis.

1. Постановка проблемы

Исследователи, занимающиеся анализом текстов в старой орфографии, особенно текстов 18–19 века, хорошо знают, что существующие лингвистические средства (спеллчекеры, лемматизаторы, программы распознавания), ориентированные на современный язык, не годятся для анализа старых текстов. Эти программы правильно анализируют формы, совпадающие с современными (*рука, новый, милость*), однако формы, отличающиеся от современных (*домъ, домъ, новаго, милостію, безсильныя, ходити*), не распознаются (считаются ошибкой), даже если все отличие сводится к конечному -ъ.

Лемматизатор mystem [1] имеет ограниченную поддержку старой орфографии путем приведения ее к современному написанию по простым формальным правилам (см. п. 3.1). Он правильно анализирует многие дореформенные написания, однако не может учесть все устаревшие формы и леммы и нередко строит для них неправдоподобные гипотезы. Для адаптации программы к языку 18–19 века необходима серьезная правка словаря и грамматических таблиц, однако код программы и словарей закрыт.

Лемматизатор Диалинг [2] и основанный на нем pymorphy [3] имеют открытый код, который можно адаптировать для старой орфографии. Однако сами словари и грамматические таблицы представлены в техническом

формате, который непрозрачен и неудобен для редактирования, а преобразование их в читаемый вид потребует серьезной переделки всего кода.

Список других морфологических анализаторов для русского языка можно найти в материалах форума «Оценка методов автоматического анализа текста» [4]. Некоторые из этих систем, помимо лемматизации, способны решать более сложные задачи: порождение гипотез для нераспознанных слов, разрешение синтаксической неоднозначности (дизамбигуация) и др. Однако все они ориентированы на современный русский язык, а возможность их настройки на другие орфографические системы и модели словоизменения представляется достаточно сомнительной.

Вместо того, чтобы пытаться переделывать чужой код, нам оказалось проще написать собственный лемматизатор, включающий возможность гибкой настройки и адаптации к языку 18–19 века, который имеет заметные орфографические, морфологические и лексические отличия от современного языка. Ключевым компонентом для решения этой задачи являются не программы и алгоритмы, а полнота и точность словаря и грамматического описания, а также удобный формат для их записи.

В принципе, для анализа текстов вместо лемматизатора можно сделать простой список словоформ с разборами. Однако, чтобы сделать такой список, нужно вначале научиться разбирать слова, причем не вручную, а автоматически. Поэтому даже на начальном этапе все равно нужен какой-то лемматизатор, в который нужно ввести хотя бы минимальную информацию о словоизменении (иначе он просто не сможет работать), а дальше обучать. Такой метод является единственно возможным для малоизученных языков, для которых нет нормального словаря и грамматического описания, поэтому они создаются индуктивно в процессе анализа существующих текстов. Однако это не относится к русским текстам 18–19 века, для которых существует обширная словарная и грамматическая база.

2. Описание программы

2.1. Определения

Морфологический анализатор (лемматизатор) — программа, выполняющая грамматический разбор текста, который включает в себя следующие задачи:

1) токенизация — разбиение текста на элементарные знаки (токены) и определение типа для каждого токена: слово, знак препинания, цифровой комплекс, тег разметки и т. д.
2) морфологический анализ для слов, присутствующих в грамматическом словаре;
3) построение гипотез для нераспознанных слов (если это возможно).

Грамматическая модель — формальное описание словоизменения данного языка, которое включает два компонента: грамматический словарь и таблица парадигм.

Грамматический словарь — 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).

Словоизменительный тип (парадигма) — набор флексий (с учетом схем чередования), общий для некоторого множества лексем. Словоизменительный тип задает соответствие между грамматическими значениями и соответствующими флексиями и чередованиями.

2.2. Принципы

Лемматизатор не привязан жестко к конкретному языку и в принципе может быть настроен на любой язык флективного типа при наличии соответствующей грамматической модели и правил токенизации. Вся конкретно-языковая информация (леммы, основы, парадигмы, флексии, граммемы) не зашита в код программы, а вынесена во внешние таблицы (словарь + парадигмы). Для записи словоизменительной информации была разработана специальная нотация, которую легко читать и править вручную.

Лемматизатор включает механизм, позволяющий ему адаптироваться к разным орфографическим системам. В процессе анализа исходный текст преобразуется во внутреннее (нормализованное) представление, которое является абстракцией от реального написания и может соответствовать нескольким орфографическим вариантам. Например, для анализа старых написаний приставок с ъ+гласная (*съиграть, съузить, съэкономить*) достаточно ввести правило замены *ъи=>ы, ъу=>у, ъэ=>э*. Механизм графических преобразований позволяет анализировать тексты с плавающей орфографией, где различные варианты написания синонимичны и свободно чередуются (*и/i/ѵ*, *ф/ѳ*, *о/ѡ*, *ъ/'*/пусто), хотя в целом анализ плавающей орфографии не сводится к простым правилам замены букв.

2.3. Алгоритм

Алгоритм работы лемматизатора состоит из следующих этапов:

1) Токенизация — программа разбивает текст на элементарные знаки (токены) и выделяет слова.
2) Нормализация — программа переводит слово во внутреннее представление, а исходная форма слова сохраняется для выдачи.

3) Программа пытается в цикле разбить слово на основу и флексию, а затем для каждого варианта разбиения проверяет дополнительные условия:
   - основа сочетается с данной флексией (имеет соответствующую парадигму);
   - основа имеет правильный вариант (при наличии чередования);
   - основа совместима с лексемой по регистру;
   - граммемы флексии совместимы с граммемами лексемы (см. п. 2.4).

   Для ускорения анализа используются заранее построенные таблицы флексий, основ и парадигм.
4) При отсутствии вариантов программа пытается построить гипотетический разбор по аналогии с существующими словами. При этом используется статистическая таблица характерных концов слова и их возможных разборов, а гипотезы сортируются в порядке убывания вероятности.

2.4. Грамматические фильтры

Правила сочетаемости граммем обычно зависят не от парадигмы, а от грамматического класса слова, и часто имеют очевидную семантическую мотивацию. Вот некоторые правила:

1) Глаголы сов. вида не имеют форм презенса (**сделающий,*сделаемый*), а формы, соответствующие презенсу несов. вида, имеют значение будущего (*делаю*=ind,pres vs. *сделаю*=ind,fut).
2) Непереходные глаголы не имеют форм пассива (*варимый, варенный* vs **веримый, *веренный*), кроме некоторых специфических конструкций (*хожено, сижено*), и не имеют медиальных форм (*-ся*), за исключением особой безличной конструкции (*не сидится, не лежится, не гуляется ему*).
3) Многократные глаголы (*сиживать*) не имеют форм презенса (**сиживаю*), а только претерита.
4) Безличные глаголы (*тошнить*) имеют только формы pres,sg,3 (*тошнит*) и past,sg,n (*тошнило*).
5) Относительные прилагательные (*вчерашний*) не имеют кратких форм и сравнительной степени.
6) Аккузатив совпадает с генитивом для одушевленных имен и с номинативом для неодушевленных, кроме некоторых парадигм (*сестру, лошадь, новую, его, их*).

Формальное наложение этих правил на парадигмы приводит к комбинаторному взрыву числа типов. Так, вместо одного типа V1 (*делать*) у нас появится 4–8–16 типов в зависимости от вида, переходности, многократности, безличности и других лексических признаков. Аналогично вместо одного типа N1 (*стол*) появится 2–4 типа в зависимости от одушевленности, лексического числа (sg/pl tantum) и др.

Механическое размножение парадигм крайне неестественно с лингвистической точки зрения и неэффективно с программной. Вместо этого в лемматизатор включены правила сочетаемости граммем, которые отсекают несовместимые комбинации или корректируют значения граммем (pres => fut).

2.5. Возможности и области применения

Созданный нами лемматизатор имеет довольно широкие возможности:

1) он умеет анализировать текст в разных орфографиях (современной, дореформенной, смешанной);
2) он умеет строить гипотезы для слов, отсутствующих в словаре;
3) он понимает различные входные форматы (включая текст с xml-подобной разметкой) и порождает несколько выходных форматов (разборы в скобках, xml-формат, табличный формат).

Лемматизатор может работать в нескольких режимах:

1) полная лемматизация, где выдаются все варианты разбора с грамматической информацией;
2) частичная лемматизация, где выдаются леммы, а грамматическая информация дается в свернутом виде или не дается вообще;
3) режим спеллчекера, где просто помечаются нераспознанные слова, но не дается грамматической информации и не порождаются гипотезы.

В настоящее время лемматизатор используется в следующих проектах:

- морфологический анализ для текстов 18 века для Национального корпуса русского языка;
- создание словарей и конкордансов к произведениям русских писателей (Ломоносов, Батюшков);
- подготовка текстов в старой орфографии для электронных библиотек;

и некоторых других.

3. Лингвистическая модель

3.1. Введение

При создании лемматизатора мы отталкивались от грамматической модели современного русского языка, зафиксированной в словаре Зализняка [5]. Мы изучили грамматические и орфографические описания 18–19 века (Ломоносов [6], Греч [7], Востоков, Грот [8]) для выявления основных особенностей языка данного периода.

Эти особенности можно условно разделить на три группы:

1) орфография (дополнительные буквы и другие правила написания);

2) словоизменение (устаревшие формы);
3) лексика (устаревшие слова).

3.2. Орфография (упрощенный вариант)

Поскольку современная орфография в основном является упрощением старой, сразу возникает идея, что старую орфографию можно анализировать путем приведения к современной. Вот основные правила преобразования старой орфографии в новую:

1) заменить устаревшие буквы на современные эквиваленты (*і=>и, ѣ=>е, ѳ=>ф, ѵ=>и*);
2) отсечь конечный -*ъ*;
3) заменить начальные *без-/в(о)з-/из-/низ-/раз-/роз-/ч(е)рез- => -с* перед глухими (NB);
4) заменить конечные *-аго/яго(+ся) => -ого/его*, *-ыя/ия(ся) => ые/ие* (NB).

Мы проверили этот метод на большом корпусе текстов в старой орфографии и получили неплохие результаты. Большая часть слов была разобрана правильно, неопознанными остались только устаревшие формы и лексемы.

Однако метод упрощения, хотя позволяет быстро получить результат, создает слишком много шума. Во-первых, механическая замена без учета морфологической структуры слова иногда работает неверно, например, слова *низкий, благо, стихия* заменяются на несуществующие формы **ниский, *блого, *стихие*. Во-вторых, смешиваются квази-омонимы, которые различаются именно буквами *ѣ–е* (*слѣзъ–слезъ, сѣлъ–селъ, свѣдѣніе–сведеніе, Вѣна–вена, морѣ–море*). В-третьих, невозможно отличить ошибочные написания, которые при переводе в современный вид совпадают с правильными (*ѣлка, ѳизика, історія*). Очевидно, что для правильного анализа старой орфографии необходим словарь и грамматическое описание, где точно указаны места употребления букв (прежде всего, *ѣ* и *ѳ*) в составе корней, суффиксов и флексий, а также учтены устаревшие формы.

3.3. Словоизменение

Если взять за основу модель словоизменения современного русского языка, то для анализа текстов в старой орфографии, помимо очевидных графических отличий, необходимо добавить в грамматическую модель формы, отсутствующие в современном языке или не учтенные в словаре Зализняка:

1) Адъективные флексии (*-аго/-яго, -ыя/-ія*).
2) Усеченные формы прилагательных (*красна/о/ы/у*), которые формально совпадают с краткими формами в именительном падеже, однако имеют и другие падежные формы (*красну*).
3) Особые формы местоимений (*ея, онъ, однъ, однъхъ*).

4) Творительный падеж 3-го склонения на *-ію* (*милостію, помощію*).
5) Сравнительная степень на *-ѣй* (*сильнѣй*) и *-яе* (*сильняе, скоряе*).
6) Вариант частицы *-ся* после гласных (*валюся, валилася*), который употребляется в современном языке (в некоторых идиолектах).
7) Деепричастия совершенного вида от основы презенса (*прийдя, увидя, взгромоздясь*), которые вполне употребительны в современном языке, но не учтены в словаре Зализняка.
8) Глагольные флексии *-ти* и *-ши* (*ходиши, ходити*), которые употребляются в основном в 18-ом веке и, видимо, должны трактоваться как церковнославянизмы.

В грамматических таблицах некоторые из этих форм имеют специальные пометы (устаревшая, церковнославянизм), чтобы их можно было отличить от общих и современных форм.

4. Лингвистические особенности текстов

4.1. Введение

Характерной особенностью языка 18–19 века является значительная лексическая и орфографическая вариативность по сравнению с современным языком. Там, где в современном языке зафиксирован один лексический вариант, в старых текстах нередко встречается несколько взаимозаменяемых вариантов. Орфографические нормы также неоднократно менялись в течение 18–20 веков (ср. орфографию Ломоносова, Греча и Грота). Авторы часто не придерживались четкой системы, издатели исправляли авторскую орфографию, а в советское время многие тексты были переизданы в модернизированной орфографии. В результате в существующих текстах наблюдается причудливое смешение разных орфографических систем, с которым вынужден работать наш лемматизатор.

Подробное описание орфографических особенностей и колебаний в истории русского языка 17–19 века содержится в работах [9], [10], [11].

Ниже мы рассмотрим некоторые языковые и орфографические особенности текстов указанного периода и возможные методы и алгоритмы их анализа. Большинство примеров взято из поэтических текстов Ломоносова, опубликованных в 8-м томе Полного собрания сочинений, где сохраняются многие особенности авторской орфографии, хотя в частично модернизированной форме.

4.2. Лексические особенности и вариативность

1) Притяжательные прилагательные.

В современном языке притяжательные прилагательные образуются в основном от собственных имен первого склонения (*Петин, Машин + мамин,*

дядин). В языке 18–19 века они образуются регулярно от всех собственных имен и многих нарицательных. Примеры: *Августов, Алцидов, Ахиллесов, Бакхусов, Варронов, Венерин, Виргилиев, Гекторов, Енеев, Дафнисов, Екатеринин, Зевесов, Ликургов, Марсов, Минервин + государев, государынин, отцов, царев*.

Сейчас лемматизатор разбирает некоторые из этих слов то как фамилии, то как косвенные формы имен. Для адекватного анализа нужно включить эти формы в расширенную парадигму имени или ввести в словарь некоторое количество притяжательных форм для обучения модуля предсказания.

2) Вариативность приставок *воз/вз*.

В современном русском языке приставки *воз/вз* четко различаются — *восстать* vs. *встать*, *восход* vs. *всход(ы)*, *воспитание* (**вспитание*), *взлететь* (**возлететь*). Лишь в некоторых однокоренных словах есть варьирование: *вспоминать–воспоминание, возлюбить–невзлюбить, вздыхать–воздыхатель*.

В языке Ломоносова эти приставки, видимо, абсолютно синонимичны и чередуются свободно. Есть десятки примеров, где между вариантами невозможно найти какого-либо смыслового или стилистического различия: *возбудить–взбудить, возвеселить–взвеселить, возвести–взвести, возгордиться–взгордиться, воздеть–вздеть, воздыхание–вздыхание, воздыхать–вздыхать, возлететь–взлететь, возложить–взложить, возлюбить–взлюбить, возмужать–взмужать, вознести–взнести, возрастить–взрастить, воспевать–вспевать, воспомянуть–вспомянуть, воспылать–вспылать, воспятить–вспятить, восток–всток* (NB)*, восточный–всточный, вострепетать–встрепетать, восход–всход, восходить–всходить*. Разумеется, это варьирование не абсолютно, некоторые слова имеют только один вариант, но все же довольно характерно.

3) Варианты суффикса *-ный/-ний*.

Примеры: *внутренний (-ый), внутренне (-о), всегдашний (-ый), дальный (-ий), дольный (-ий), искренний (-ый), крайний (-ый), осенний (-ый)*. В современном языке такое колебание сохранилось в слове *междугородний (-ый)* и в паре *искренний–искренно*.

4а) Варьирование *и/ъ* в конце основы.

Примеры: *видение (-ье), владение (-ье), возмущение (-ье), внимание (-ье), волнение (-ье), дыхание (-ье), желание (-ье), житие (-ье), здравие (-ье), копье (-ие), оружие (-ье), щастие (щастье)*.

В современных словарях обычно зафиксирован один вариант: *варенье* vs. *варение*, *беганье* vs. *плавание*. Однако в реальности такие варианты возможны почти для любого слова и воспринимаются как свободные колебания типа *водой–водою*. Эти варианты активно используются в поэтической речи для подстановки слова в различные ритмические контексты.

Для адекватного анализа таких вариантов мы добавили в словарь и в программу условный символ (гиперграфему), который может соответствовать обоим вариантам — *и/ъ*.

4б) Варьирование *и/ъ*+гласная в середине основы.

Примеры: *азиатец–азиятец, италианский–италиянский–итальянский, баталион–батальон, материал–матерьял, вариант–варьянт*.

Это варьирование аналогично предыдущему, но лексически более ограничено, хотя некоторые колебания сохраняются и в современном языке. Оно похоже на характерное соотношение между церковнославянскими и русскими формами: *бию–бью, Татиана–Татьяна, Иулиан–Юлиан–Ульян*.

4.3. Орфографические особенности и вариативность

Основные точки вариативности в старой орфографии — те же, что и в современной:

1) обозначение мягкости шипящих (*меч* vs. *ночь*, *куш* vs. *идешь*) и др. согласных (*росте* vs. *бросьте*);
2) обозначение *ё/о* после шипящих и *ц* (*чёрт* vs. *чохом*, *шёлк* vs. *мешок*);
3) глухие/звонкие согласные в позиции нейтрализации (*воздать* vs. *воскресить*);
4) употребление прописные букв (*Англия* vs. *английский, англичанин*);
5) слитное/раздельное/дефисное написание (*тоже* vs. *то же*, *надвое* vs. *по двое*, *вмиг* vs. *в минуту*).

В истории русской орфографии правила написания в этих точках неоднократно менялись, а для случая 5 (слитно/раздельно) внятные правила не выработаны до сих пор.

Вот некоторые примеры орфографических колебаний в текстах Ломоносова и способы их решения:

1а) Написание ь/ноль после мягких согласных в основе:

вер(ь)вь, вер(ь)х, вет(ь)вь, пер(ь)вый, цер(ь)ковь, гор(ь)кий, сил(ь)ный, т(ь)ма, воз(ь)му.

Это в основном лексические колебания, касающиеся конкретных слов, но вообще факультативность написания ь внутри слова характерна для более раннего периода русского письма.

1б) Написание ь/ноль после шипящих во флексиях:

будеш(ь), видиш(ь), возбудиш(ь), восстаеш(ь), даеш(ь), держиш(ь), дерзнеш(ь), желаеш(ь), знаеш(ь), зриш(ь), идеш(ь), имееш(ь), любиш(ь), боиш(ь)ся, возмутиш(ь)ся.

бичь, ключь, лучь, мечь, Васильевичь, Ивановичь, Михайловичь;

Для анализа таких форм мы добавили флексии *-ш* и *-шся* в парадигму глагола, а также флексию -ь для имен на ч.

2) Написание *е/о* после шипящих и *ц*:

агньцов, венцев–венцов, венцем, зайцов, концев–концов, лисицой–лисицей, лицо–лице, лицом–лицем, любимцов, месяцов–месяцев, младенцов,

отцев–отцов, отцем–отцом, праотцов, пришельцов, пшеницой–царицей, свинцом–певцем, стрельцов–пловцев;

несчотный–несчетный, предпочол–предпочел, чорной–черной, чорны–черны, мужичок, сверьчок, старичок, плечо, счота, учоный, чорт — бичем;

большой–большей, большом–большем, шол–шел, вшол–вошел, обшол–обшел, отшол–отшел; поджог, рожок — душей, душею; чужом–чужем, чужою–чужею.

Колебание во флексиях можно решить алгоритмически, если ввести оба флексий (*о/е, ом/ем, ов/ев*) во все соответствующие парадигмы, независимо от места ударения. Колебание в основе можно решить, только если ввести «гиперграфему» *о/е* в соответствующие лексемы, однако в любом случае это будет словарное правило.

3) Прилагательные на *-зкий* (аналогия с частотными на *-ский — росский*): *близкий–блиский, дерзкий–дерзский–дерсский, мерзкий–мерзский–мерсский*.

Это индивидуальная особенность данных слов, которая не решается алгоритмически.

4) Прописная буква используется не только в собственных именах, но также в следующих случаях:

титулы — *Император, Самодержец, Величество + Генерал Порутчик, Академик, Кавалер* (иногда);

названия народов (не всегда) — *Немцы, Немецкий, Шведы, Шведский*;

производные от топонимов — *Азийский, Американский, Афинский, Багдатский, Берлинский*;

названия месяцев, наук, искусств, олицетворения и др.

Прописная буква не составляет проблемы для лемматизатора, поскольку в русском языке любое слово, нормально пишущееся со строчной, получает первую прописную в определенных контекстах. Однако это небезразлично для словаря, где нужно выбрать словарную форму слова, исходя из наиболее частотного написания, а также для спеллчекера, где регистр важен для проверки правильности слова.

5) Слитно или раздельно:

не+глаголы, наречия и др. — *небойся, недай, недайте, неокончав, неопасайся, неиначе, неможно*;

предлоги (редко) — *длятого, изовсего, вних*;

частицы *б(ы), ж(е), ли(ль)* — *былаб(ы), вернаяб, весьмабы, взойтиб, возможноли(-ль), вамиб, гдеб, гдеж, горыль, давноб, давноль, довольноль, долголь, Елисаветаб, естьлиб(ы), молвиж, моглиб, могуль, намиб, мыль, нашаб, нуж, ониб.*

Для анализа таких форм мы включили в программу дополнительное правило для слитных клитик (*не, бы, же, ли*), которое срабатывает, если прямой анализ не дал результата.

6) Написания *сч/щ, жч/щ*:

счастье–щастье, счастливый–щастливый, разсчет–расчет–ращет, мужчина–мущина.

Это индивидуальная особенность данных слов, которая не решается алгоритмически.

В целом, для анализа орфографической вариативности нет универсального метода, но можно предложить частные решения для конкретных случаев:

1) Нормализация — преобразование исходного написания в нормализованное, которое далее анализируется по стандартной грамматической модели.
2) Расширение грамматической модели — включение дополнительных форм в парадигмы.
3) Расширение лексической модели — специальная нотация для записи лексической вариативности («гиперграфемы») или включение дополнительных вариантов в словарь.
4) Модификация алгоритма для анализа слитных написаний.

5. Заключение

Для анализа текстов 18–19 века необходимо создание полноценной словарной базы, прежде всего, **грамматического словаря**, который содержит все основные лексемы и грамматические формы, характерные для языка данного периода. В качестве лексической базы для такого словаря нужно использовать статистические данные, полученные из реальных текстов, а также следующие лексикографические источники:

- Словарь Академии Российской (1789–1794);
- Словарь церковнославянского и русского языка (1847);
- Полный русский орфографический словарь (1898);
- Словарь русского языка 18 века;

и некоторые другие.

Литература

1. *Морфологический* анализатор mystem (http://company.yandex.ru/technologies/mystem/).
2. *Сокирко А. В.* Русский морфологический словарь (http://aot.ru/docs/rusmorph.html).
3. *Морфологический* анализатор Pymorphy (http://bitbucket.org/kmike/pymorphy/).
4. *Оценка* методов автоматического анализа текста: морфологические парсеры русского языка (http://ru-eval.ru/participants.html).
5. *Зализняк А. А.* Грамматический словарь русского языка. — М., 2008.

6. *Ломоносов М. В.* Российская грамматика. — СПб., 1788.
7. *Греч Н. И.* Практическая грамматика русского языка. — СПб., 1827.
8. *Грот Я. К.* Спорные вопросы русского правописания от Петра Великого доныне. — СПб., 1873.
9. *Ильинская И. С.* Лексика стихотворной речи Пушкина. «Высокие» и поэтические славянизмы. — М., 1970.
10. *Перцов Н. В.* О соотношении письменной и устной форм поэтического языка (К вопросу о функциональной нагруженности старого русского правописания) // Вопросы языкознания. 2008. № 2. С. 30–56.
11. *Каверина В. В.* Становление русской орфографии в XVII–XIX вв.: правописный узус и кодификация. Дисс. ... докт. фил. наук. — М., 2010.

References

1. *Mystem* morphological analyzer (http://company.yandex.ru/technologies/mystem/).
2. *Sokirko A. V.* Russian morphological dictionary (http://aot.ru/docs/rusmorph.html).
3. *Pymorphy* morphological analyzer (http://bitbucket.org/kmike/pymorphy/).
4. *Natural* language processing evaluation: morphological parsers of Russian (2010) (http://ru-eval.ru/participants.html).
5. *Zaliznjak A. A.* (2008) Grammaticheskij slovar' russkogo jazyka [Russian grammatical dictionary]. Moscow.
6. *Lomonosov M. V.* (1788) Rossijskaja grammatika [Russian grammar]. St.-Petersburg.
7. *Grech N. I.* (1827) Prakticheskaja grammatika russkogo jazyka [Practical grammar of the Russian language], St. Petersburg.
8. *Grot Ja. K.* (1873) Spornye voprosy russkogo pravopisanija ot Petra Velikogo donyne [Controversial issues in the Russian spelling from Peter the Great till now]. St.-Petersburg.
9. *Il'inskaja I. S.* (1970) Leksika stihotvornoj rechi Pushkina. "Vysokie" i poèticheskie slavjanizmy [The Vocabulary of Pushkin's verse speech: "Sublime" and poetic slavonicisms], Nauka, Moscow.
10. *Pertsov N. V.* (2008) On the interrelation between the written and oral forms of the poetic language (On the issue of functional capacity of the pre-reform Russian spelling) [O sootnoshenii pis'mennoj i ustnoj form poèticheskogo jazyka (K voprosu o funkcional'noj nagruzhennosti starogo russkogo pravopisanija)], Voprosy jazykoznanija (Issues in Linguistics), no. 2. pp. 30–56.
11. *Kaverina V. V.* (2010) Stanovlenie russkoj orfografii v XVII–XIX vv.: pravopisnyj uzus i kodifikacija [Formation of Russian spelling in the 17–19th centuries: usage and codification]. D. Phil. thesis. Moscow.