

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

Информационные технологии и письменное наследие

El'Manuscript-2012

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

Петрозаводск, Ижевск
2012

УДК 004.9
ББК 81.11+81.2-0
И741

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

Ответственные редакторы:

В. А. Баранов, д-р филол. наук, проф.

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

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

ISBN 978-5-8021-1402-5

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

УДК 004.9
ББК 81.11+81.2-0

ISBN 978-5-8021-1402-5

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

ЛЕММАТИЗАТОР ДЛЯ ДОРЕФОРМЕННОЙ РУССКОЙ ОРФОГРАФИИ¹

А. Е. Поляков

Научная педагогическая библиотека им. К. Д. Ушинского, Москва

Current linguistic processors are not suitable for analysis of pre-reform Russian texts due to numerous graphical, morphological and lexical differences from the modern language. We have developed a lemmatizer which can properly analyze pre-reform texts and has a facility for flexible adaptation to other spelling systems.

1. Постановка проблемы

Существующие лингвистические процессоры для русского языка (спеллчекеры, лемматизаторы, программы распознавания) непригодны для анализа текстов в старой орфографии. Во-первых, обычно они не воспринимают дополнительные буквы (Ѣ, ѣ, ѿ, ѹ) и другие орфографические особенности, даже самые тривиальные, типа конечного -ѣ. Во-вторых, они ориентированы на современный русский язык и модель словоизменения. Между тем, язык 18–19 века имеет заметные морфологические и лексические отличия от современного языка. Некоторые программы включают ограниченный анализ старой орфографии путем приведения ее к современной (см. п. 3.1), однако они не могут учесть устаревшие формы и леммы. Исходный код этих программ обычно закрыт, а словари представлены во внутреннем формате, который непрозрачен и неудобен для редактирования, поэтому мы были вынуждены разрабатывать собственный лемматизатор.

2. Описание программы

2.1. Определения

Лемматизатор — программа, выполняющая грамматический разбор текста, который включает в себя следующие задачи:

- 1) токенизация — разбиение текста на элементарные знаки (токены) и определение их типа (слово, знак препинания, число, тег разметки и т. д.);
- 2) морфологический анализ для слов, имеющих в словаре;
- 3) построение гипотез для нераспознанных слов (если это возможно);
- 4) анализ многословных сочетаний (если они есть в словаре).

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

Грамматический словарь — список лексем с приписанной информацией о словоизменении, которая включает, как минимум, следующие параметры:

- 1) основа с указанием чередований;
- 2) постоянные признаки лексемы (часть речи, род, одушевленность...);
- 3) код словоизменительного типа (парадигмы). **Словоизменительный тип (парадигма)** — набор флексий, общий для

некоторого множества лексем, который задает соответствие между грамматическими значениями и соответствующими им формами.

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

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, грамлеммы) не зашита в код программы, а вынесена во внешние таблицы. 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. Этот механизм позволяет анализировать тексты с плавающей орфографией, где почти любое слово может иметь несколько вариантов написания (и/и', ф/ф', ь/ы, ь'/пусто).

2.3. Алгоритм

Алгоритм работы лемматизатора состоит из следующих этапов:

- 1) Токенизация – программа разбивает текст на 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 нужную парадигму);
 - основа имеет правильный вариант (при наличии чередования);
 - основа совместима с лексемой по регистру;
 - грамлеммы флексии совместимы с грамлеммами лексемы.

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. Возможности и области применения

Разработанный нами лемматизатор имеет довольно широкие возможности:

- 1) он умеет анализировать текст в разных орфографиях;
- 2) он умеет строить гипотезы для слов, отсутствующих в словаре;
- 3) он понимает несколько входных форматов, включая текст с xml-подобной разметкой, и порождает несколько выходных форматов.

Программа может работать в следующих 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 грамматическая информация и не порождаются гипотезы.

В настоящее время лемматизатор используется в следующих проектах:

- Создание конкорданса к текстам М.В.Ломоносова.
- Создание корпуса текстов и словаря языка К.Н. Батюшкова.
- 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. Далее, мы проанализировали основные особенности языка XVIII–XIX вв. на основе грамматических описаний и реальных текстов. Эти особенности можно условно разделить на три группы:

- 1) орфография (дополнительные буквы, другие правила написания);
- 2) словоизменение (устаревшие формы);
- 3) лексика (устаревшие слова).

3.1. Орфография (упрощенный вариант).

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

- 1) заменить устаревшие буквы на современные (і=>и, ѣ=>е, ѳ=>ф, ѵ=>и);
- 2) отсечь конечный -ѣ;
- 3) заменить без-/во)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 и получили неплохие результаты. Большая часть слов была разобрана правильно, неопознанными остались устаревшие формы и лексемы.

Однако метод упрощения, хотя позволяет быстро получить результат, дает слишком много шума. Во-первых, механическая замена букв без учета морфологической структуры слова иногда работает неверно (правила 3 и 4). Во-вторых, смешиваются слова, которые различаются буквами ѣ–е (слѣзь – слѣзь, сѣль – сѣль, свѣдѣніе – сведеніе, Вѣна – вена, морѣ – море). В-третьих, невозможно отличить ошибочные написания, которые при переводе в современный вид совпадают с правильными. Очевидно, что для полноценного 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.

3.2. Словоизменение

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

- 1) Адъективные флексии (-аго/-яго, -ья/-ія).
- 2) Усеченные формы прилагательных (красна/о/ы/у), которые формаль-но совпадают с краткими формами, однако имеют и другие падежи (красну).
- 3) Особые формы местоимений (ея, онѣ, однѣ, однѣхъ).
- 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 также в современном языке (в некоторых идиолектах)
- 7) Деепричастия совершенного вида от основы презенса (приидя, увидя, взгромоздясь), которые вполне употребительны в современном языке.
- 8) Глагольные флексии -ти и -ши (ходиши, ходити), которые характерны для 18-го века 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.

3.3. Орфографическая вариативность

Характерной особенностью языка XVIII–XIX вв. по сравнению с современным является значительная лексическая и орфографическая вариативность. Там, где в современном языке зафиксирован один вариант слова, в старых текстах можно встретить несколько взаимозаменяемых вариантов. Орфографические правила неоднократно менялись в течение XVIII–XX вв., а в советское время многие тексты были переизданы в частично модернизированной орфографии. Поэтому в реальных текстах наблюдается причудливое смешение разных орфографий, с которым вынужден работать наш лемматизатор.

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

- 1) обозначение мягкости для шипящих (мечь — меч, идешь — идеш, ночьной — ночной);
- 2) обозначение ё/о после шипящих и ц (чёрный — чорный, дьячѣк — дьячок, шѣл — шол, лицо — лице, зайцов — зайцев);
- 3) глухие/звонкие согласные в позиции нейтрализации (возсиять — воссиять, восток — возток, дерзкий — дерзский — дерсский — дерский);
- 4) употребление прописных букв (Английский, Англичане, Славенский);
- 5) слитное/раздельное/дефисное написание (кто нибудь, повидимому, по русски, до тла, вшутку, то-есть, кто-бы — ктоб, небойся, долголь);
- 6) отражение фонетических колебаний, например и/й/ь (дыхание — дыханье, счастье — счастье, вариант — варьянт, итальянский — итальянский).

Для анализа орфографической вариативности нет универсального метода, но можно предложить решения для конкретных случаев :

- 1) Нормализация — преобразование написания в нормализованную форму, которая 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.

4. Направления дальнейшей работы

Для анализа текстов XVIII–XIX вв. 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 XVIII века
- и другие словари.