

Организация поиска в текстовых коллекциях на русском языке XVIII века

Соловьёв В.Д.
КГУ, ИПИ АН РТ
MAKI.solovyev@mail.ru

Маргулис И.С.
ИПИ АН РТ, ЦИТ КГУ
ilya_margulis@mail.ru

Аннотация

Исследование старорусских текстов существенно облегчается при использовании их электронных версий. Материалы становятся доступными для исследователей не только из крупных университетских центров. Оцифровка старинных текстов ведётся крупнейшими библиотеками Мира, в том числе Российской государственной библиотекой. Актуальность и различные аспекты проблемы оцифровки недавно были подробно обсуждены в [41].

Исследование старорусских текстов в первую очередь требует разработки морфологических анализаторов и создания поисковых программ, ориентированных на словоизменение в старорусском языке. Для решения проблемы наиболее перспективной выглядит идея адаптации современных алгоритмов морфологического анализа и поиска к старорусскому языку с использованием электронного словаря XVIII века.

В ходе выполнения проекта впервые созданы прямой и обратный электронные словари XVIII века. Разработана программа для локального поиска слов в старинной орфографии с учётом словоизменения для русского языка XVIII века. Программа представляет собой инсталлируемую надстройку для MS Word. В настоящее время при поддержке РФФИ и РГНФ¹ создаются корпуса текстов XVIII века, в которых могут быть использованы результаты данной работы.

1. Введение

1.1 Цели исследования

Целью исследования является создание программных средств для работы с текстами на старорусском языке в оригинальной орфографии. В перспективе это способствует развитию эволюционной и исторической лингвистики (в первую очередь относительно русского языка XVIII века), корпусных методов исследований.

В основе задачи создания поисковых средств для работы с текстами в старинной орфографии лежит три основных проблемы: кодировка, морфологический анализ и интерфейс.

1.2 Обзор специализированных программных средств для работы с русскоязычными документами в старинной орфографии; проблема кодировок

Существуют четыре основных известных программных средства для работы с текстами на русском языке в старинной орфографии. Их представляют этимологический проект *STARLing*, проект по разработке шрифтов «Ирмологий»,

информационно-поисковая система (ИПС) «Манускрипт» и древнеязычный текстовый процессор «Книжица». Ниже данные разработки рассмотрены более подробно.

Программа *STARLing* [32] представляет собой лингвистическую СУБД и в качестве основного предназначения имеет работу с многоязычными этимологическими базами данных. Последняя версия программы (*Star4win v. 2.00, September 2004*) использует кодировку Unicode и имеет возможность работы с символами русского языка XVIII века, однако в целом она не ориентирована на старорусский язык. *STARLing* решает такие задачи, как преобразование письменного текста в многоуровневую текстовую базу данных (ТБД); разметка ТБД (автоматическая и полуавтоматическая с интерредактированием); создание и корректировка словарных БД (с опорой на внешние источники).

Программа *STARLing* имеет возможность конвертирования файлов в форматы RTF и HTML, но не непосредственной работы с данными файлами. *STARLing* требует освоения специфического интерфейса и некоторых навыков программирования для широкого использования программы.

Три других проекта ориентированы на древнерусский язык (до XVIII века). В кодировке Unicode, являющейся в настоящее время стандартом де-факто, отсутствуют многие символы древнерусского языка. Это заставляет разработчиков прибегать к использованию нестандартных кодировок, что вызывает проблемы совместимости.

Программа «Ирмологий» [11] позволяет создавать и редактировать древнерусские тексты. Она реализована в качестве надстройки-расширения для MS Word. Программа «Ирмологий» базируется на нестандартной мультишрифтовой кодировочной системе *Irmologion* (использующей до 13 шрифтов), что приводит к несовместимости со стандартными кодировками. Встроена возможность конвертирования только в две альтернативные кодировки для старорусского языка: UCS8 и HIP. Интерфейс требует частого переключения с одного шрифта на другой.

ИПС «Манускрипт» [17] также ориентирована на работу с более древними текстами и использует собственную мультишрифтовую кодировочную систему. Важным компонентом ИПС «Манускрипт» является специально созданный для него редактор

древних текстов *OldEd*. В нём отсутствует возможность непосредственной работы с файлами распространённых форматов. Программа требует освоения специфического интерфейса.

Древнеязычный текстовый процессор «Книжица» также использует собственный формат данных. Он несовместим с другими кодировками, возможность конвертирования отсутствует. В связи с этим данный самостоятельный проект был приостановлен, программа не модернизировалась, сайт перестал обновляться в 1999 году и был закрыт.

Не смотря на то, что множество символов, используемое в альтернативных кодировках для древнерусского языка (до XVIII века), достаточно для печати текстов XVIII века в оригинальной орфографии, использовать альтернативные кодировки для печати текстов XVIII века не целесообразно по двум причинам. Во-первых, это приводит к несовместимости кодировок. Во-вторых, не исключено дальнейшее расширение кодировки Unicode для древнерусского и древнеславянских языков, поэтому более вероятно, что при создании электронных корпусов старинных текстов древнерусские тексты, набранные в альтернативных кодировках, придётся конвертировать в Unicode, чем старорусские (XVIII-XIX век) – в какую-либо нестандартную кодировку для древнерусских текстов. В рамках коллектива ранее была произведена работа над созданием принципов пополнения Unicode для древнерусских и древнеславянских азбук [18]. Дальнейшим продвижением проекта занимается ассоциация «Электронные издания/представления славянских рукописей» [31], которое именуется также сообществом «Письменное наследие».

Особенности старинных документов (изменение цвета бумаги, пятна, механические повреждения листов, риск порчи документов при сканировании, выцветание красок) вызывают затруднения для распознавания и создания текстовых коллекций в старорусской орфографии. Однако в настоящее время Интернет регулярно пополняется такими текстами [10,9,22], решаются проблемы распознавания [38]. В рамках коллектива ранее были созданы макросы для частичной автоматической коррекции текстов в оригинальной орфографии XVIII века, распознанных при помощи программы *ABBYY FineReader* алгоритмами распознавания с обучением и без обучения. Это упрощает создание коллекций и корпусов текстов. Также коллективом была произведена классификация ошибок распознавания, позволяющая корректно подбирать слова для замены в некачественно распознанных текстах, и были установлены статистические особенности, тактически важные для повышения качества распознавания старорусских текстов.

Можно предположить, что со временем многие сохранившиеся старинные документы будут

представлены в электронном виде в текстовом формате для исследований. На филологических сайтах (например [39,21,28]), форумах (например, [14]) и сайтах по культуроведению (например, [37]), а также в некоторых энциклопедических статьях (принадлежащих, например, энциклопедии «Википедия» [36]) многие цитаты приводятся в старинной орфографии. Кроме того, в сети Интернет имеется ряд современных текстов с элементами старинной орфографии. К ним относятся современные церковные тексты (православные, старообрядческие, нетрадиционные), публикации коммерческих издательств, специализирующихся на религиозной литературе (например, [8]), ряд философских, сатирических, публицистических и других произведений. Элементы орфографии XVIII века используются как внутренний язык в интернет-сообществах и блогах, посвящённых исторической и литературной тематике, филологии. Специалисты и любители заинтересованы в использовании качественного поиска документов в орфографии XVIII века, данная задача является в настоящее время актуальной [10,9].

Для текстов XVIII века использование кодировки Unicode теряет смысл лишь в тех случаях, когда текст набирается исключительно для конкретного печатного, неэлектронного издания – тогда ни кодировка, ни формат не имеют значения. В этом случае печать может производиться даже из графического файла, однако для качественных (статистических и прочих) исследований с применением компьютерных технологий необходимо использовать текстовый формат данных.

Таким образом, выбор кодировки для работы с текстами XVIII века определённо останавливается на Unicode.

1.3 Поиск для русскоязычных текстов в оригинальной орфографии XVIII века

1.3.1 Словари

Существенную роль в организации поиска играют словари. За 2001-2006 гг. были переизданы в печатном виде все тома «Словаря Академии Российской» [2,30] (САР) 1789-1794 гг. Данная современная версия САР представляет собой наборное издание факсимильного типа (не путать с факсимильным изданием). Её электронный вариант [33] содержит также факсимильную версию САР, но он не распространяется свободно в сети Интернет.

1.3.2 Настольные приложения

Многопрофильные настольные текстовые редакторы и процессоры, в достаточной мере поддерживающие Unicode (например, *MS Word*, *MS Notepad*), позволяют осуществлять поиск со старорусскими символами без учёта словоизменения. Для *OS MS Windows* и её Unicode-поддерживающих приложений единственным

известным средством для непосредственной печати старорусских символов в поисковой строке (без копирования и вставки) является разработанная ранее одним из авторов клавиатурная раскладка [19]. Она позволяет осуществлять поиск на диске без учёта словоизменения при помощи стандартных средств *OS MS Windows*, в которых имеется полная поддержка Unicode (*Windows XP* и выше).

Специализированные настольные программы для работы с русскоязычными текстами в старинной орфографии (*STARLing*, «Ирмологий», «Книжица»), описанные в предыдущем подразделе, не используют морфологический анализ в своих функциях, в том числе в функции поиска, если она имеется.

1.3.3 Web-приложения

На данный момент поиск русскоязычных текстов в орфографии XVIII века в Интернете возможен в таких поисковых машинах, как *Google*, *MSN*, *Yahoo*, *AOL (Netscape)*, *USSeek*, *AltaVista*, *Alltheweb*, *ASK*, *Scirus*, *Nigma* и расширенный *Яндекс*, однако в них поиск производится только по исходным формам набранных пользователем ключевых слов без учёта морфологии, а также требует использовать дополнительные средства для набора символов XVIII века. В качестве таких средств предлагается, например, виртуальная клавиатура, размещённая на странице проекта «Русский поиск» [25] (данная ссылка может оказаться недействительной, так как сайт периодически меняет URL), позволяющая печатать старорусские буквы в поисковой строке *Google*. Также, в качестве более универсального средства можно использовать упомянутую выше клавиатурную раскладку, позволяющую печатать все старорусские буквы в строке запросов любой поисковой машины. Последняя разработка в отличие от виртуальной клавиатуры «Русский поиск» позволяет производить поиск в канонически набранных старорусских текстах, где все используемые буквы принадлежат кириллическому диапазону Unicode. Это не исключает возможности поиска и в текстах, набранных не канонически, раскладка легко переключается. Под не канонически набранными текстами XVIII века здесь подразумеваются тексты, в которых, например, вместо букв «И-десятеричное» и «Ижица» из кириллического диапазона Unicode используется графически похожие на них буквы «И-десятеричное» и «V» из латинского диапазона, вместо буквы «Фита» – «Тэта» из диапазона греческих символов и т.п. Расположение символов в созданной раскладке рассчитано по эргономической модели соответственно частотам двух- и трёхбуквенных сочетаний в старорусских текстах. Раскладка представляет собой устанавливаемый файл локали *Windows* и используется специалистами по истории языка и языкознанию. Раскладка позволяет также печатать старорусские символы для поиска непосредственно в открытом через браузер или

редактор документе, используя стандартные средства поиска.

Основным препятствием для внедрения поиска тестов XVIII века в поисковых системах является отсутствие полной поддержки поисковыми системами кодировки Unicode. В поисковых машинах *Anopm*, *Metabot*, облегчённый *Яндекс*, *Rambler* и в большинстве ftp-поисковиков некоторые Unicode-символы заменяются символами однобайтовых кодировок по принципу наложения кодовых страниц Unicode, либо двухбайтный код обрабатывается как два символа (например, в метапоисковой системе *Metabot.ru*), либо генерируется ошибка. По принципу наложения кодовых страниц дополнительные Unicode-символы заменяются в ряде случаев графически, лексически и/или этимологически похожими символами однобайтовых кодировок, а в ряде случаев заменяются никаким образом не похожими символами однобайтовых кодировок. Например греческая буква Пси (Ψ) заменяется кириллической буквы Ша и т.п. В поисковых машинах *Google*, *MSN*, *Yahoo*, *AOL (Netscape)*, *USSeek*, *AltaVista*, *Alltheweb*, *ASK*, *Scirus*, *Nigma* и расширенный *Яндекс* кодировка Unicode поддерживается более широко, в том числе возможно использование Unicode-символов русского языка XVIII века.

Таким образом, можно выделить 4 уровня способности поисковых систем обрабатывать запросы, содержащие символы, отсутствующие в однобайтовых кодировках, в том числе символы старорусского языка (в качестве примеров приведены популярные в России поисковики):

0. Морфологическое расширение поиска для старорусского языка не реализовано в поисковых системах. Если слово запроса состоит исключительно из букв современного русского языка, поиск производится с учётом морфологии современного русского языка, которая не совпадает с морфологией старорусского языка.

1. Поисковые машины *Google*, *MSN*, *Yahoo*, *AOL (Netscape)*, *USSeek*, *AltaVista*, *Alltheweb*, *ASK*, *Scirus*, метапоисковый стартап *Nigma* и расширенный поиск *Яндекс* представляют базовые возможности поиска выражений, включающих Unicode-символы, отсутствующие в однобайтовых кодировках, в том числе символы старорусского языка. По причине использования не Unicode-полного шрифта (реже, например, в *Nigma* – по другим причинам) отображение результатов поиска в браузере может быть некорректным, но поиск является точным.

2. Web-поисковики *Anopm*, *Metabot* (метапоисковик), облегчённый поиск *Яндекс*, *Rambler* (и расширенный поиск *Rambler*), большинство известных ftp-поисковиков (с посттранслитерацией) подвергают некоторые Unicode-символы запроса унификации, что приводит к некорректному поиску, либо генерируют ошибку. Унификация заключается в замене некоторых дополнительных Unicode-

символов на (графически, лексически и/или этимологически) похожие или не похожие символы однобайтовых кодировок, например, по принципу наложения кодовых страниц Unicode. Например, для слов с буквами старорусского языка облегченный *Яндекс* и *Rambler* дают ошибку, *Апорт* и *Metabot* – ошибку или некорректное отображение.

3. Поисковые машины *Mail.ru*, *Webalta*, *Bigmir.net*, *Lupa.ru*, и метапоисковый стартап *Punto.ru* дают ошибки для выражений, содержащих символы, отсутствующие в однобайтовых кодировках. *Lupa.ru* использует транслитерацию для ряда символов (например, кириллические символы национальных алфавитов заменяются графически похожими буквами русского алфавита).

Пункты 1-3 текущего раздела отражают специфику работы поисковых систем с кодами символов, а гипотетический пункт 0 – морфологическое расширение поиска для старорусского языка. Модернизация поисковых машин, переход с уровней 2 и 3 на уровень 1 является возможным.

Специализирующийся на древнерусских текстах ИПС «Манускрипт» использует анализ по морфологии для древнерусского языка (до XVIII века), он это не представляет возможностей для поиска с учётом словоизменения старорусских текстов (XVIII-XIX век), так как данные морфологии отличаются. База данных ИПС «Манускрипт» содержит несколько текстов XVIII века (10 писем М.В.Ломоносова), однако они приведены в современной орфографии.

1.3.4 Итоги

Растущие объёмы документов, использующих символы, размещённые на кириллической странице Unicode, но не принадлежащие современному русскому алфавиту, нуждаются в таких механизмах работы, как поиск с учётом словоизменения. Это относится и к текстам в орфографии старорусского языка.

В настоящее время морфологический анализ и поиск словоформ для языка XVIII века не реализован ни в одном широкопрофильном или специализированном, настольном или Web-ориентированном программном комплексе (за исключением представленной ниже разработки). Библиография на тему автоматизированного поиска в текстах XVIII века с учётом морфологических изменений отсутствует.

С большой вероятностью можно предположить, что все крупные поисковые системы со временем введут более полную поддержку Unicode. Это предоставит возможность реализовать морфологическое расширение, например, для поиска в текстах в оригинальной орфографии старорусского языка.

Создание новых специальных поисковых машин, в том числе для старорусского языка, не является целесообразным в связи перспективой скорого внедрения морфологических расширений в

развитых поисковых системах (а также в специализированных лингвистических комплексах).

1.4 Вопросы интерфейса

В качестве основного принципа организации интерфейса для поиска в текстах на старорусском языке целесообразно выбрать максимальную интеграцию со стандартными средствами работы с текстовыми документами, в том числе поддержку распространённых форматов. Предпосылки для этого были изложены в пунктах 1.2 и 1.3.

2. Описание исследования

2.1 Идея исследования

Исследование заключается в разработке морфологического расширения для поиска русскоязычных текстов XVIII века, в создании необходимых словарей и оценке эффективности предложенных алгоритмов.

2.2 Создание словарей

При участии сотрудников библиотеки им. Лобачевского КГУ осуществлён ручной набор слов из САР (Словарь Академии Российской, включает более 42000 слов). Набранные 6 томов объединены в прямой и обратный словники. Для данного словаря одно старорусское слово содержит в среднем 9 букв.

Разработана программа для интерактивного пополнения словаря: создание словарей из старорусских текстов и интеграция в расширенный прямой или обратный САР. Программа работает с любыми документами, доступными для просмотра через *MS Word*, в том числе с pdf-файлами, если установлена соответствующая надстройка.

Совместно с заведующим Кафедрой истории русского языка и языкознания (ИРЯЯ) КГУ Николоаевым Г.А. создан словарь псевдоокончаний старорусского языка. Текущая версия словаря включает 136 псевдоокончаний (включая нулевое окончание), в среднем 2,7 буквы в одном псевдоокончании.

В качестве словаря псевдооснов было использовано два словаря. Первый создан на основе морфологической таблицы современного русского языка, используемой комплексом программ «Рабочее место лингвиста» (РМЛ, она же система «Диалинг») [24]. Второй словарь псевдооснов создан из САР методом стемминга для псевдоокончаний старорусского языка. Планируется коррекция данного словаря, после которой его можно будет эффективно использовать вместо современного словаря.

В дальнейшем планируется произвести морфологическую разметку словаря псевдооснов на базе САР. Подробная морфологическая разметка старорусского языка необходима для развития эволюционной и исторической лингвистики, однако не все грамматические характеристики могут найти

применение в алгоритмах поиска. Высокая вариативность старорусского языка осложняет определение безыключительных флективных классов и ставит под вопрос целесообразность их использования в поисковых алгоритмах. С большой вероятностью это приведёт к снижению объёма выдачи за счёт исключения результатов, релевантных запросу. Однако определение парадигм чередований псевдооснов в лексемах необходимо для повышения качества поиска в старорусских текстах.

2.3 Алгоритм поиска

Исследование старорусского языка показало, что для организации поиска целесообразно использовать алгоритм, близкий к алгоритму, применяемому для поиска текстов на современном русском языке [40]. В общих словах, последовательность операций включает в себя стемминг, поиск словарной псевдоосновы и генерацию словоформ.

Морфологическая разметка словарей требует времени. В качестве оптимального алгоритма поиска на базе имеющихся в настоящее время словарей представляется следующий алгоритм:

1. Стемминг: в слове запроса производится выделение всех возможных морфологических псевдооснов, пользуясь списком псевдоокончаний.

2. Для каждой выделенной таким образом потенциальной псевдоосновы определяется её присутствие в словаре псевдооснов.

3. Для каждой псевдоосновы, если она оказалась словарной, поиск будет осуществлен по всем её словоформам, образуемым от псевдоосновы и псевдоокончаний.

4. Если ни одна потенциальная псевдооснова не оказалась словарной, то пользователю-лингвисту предлагается в специальной форме выделить подсветкой псевдооснову слова запроса и добавить её в словарь и то же самое сделать для псевдоокончания. Пользуясь данной формой, пользователь может напечатать и добавить практически неограниченное количество псевдооснов и псевдоокончаний.

5. Если ни одна потенциальная псевдооснова не оказалась словарной, и пользователь не произвёл действий, описанных в предыдущем пункте, то поиск производится на базе потенциальных псевдооснов, выделенных благодаря списку известных псевдоокончаний. В наихудшем случае это одно нулевое окончание.

6. Процесс поиска позволяет пользователю-лингвисту убедиться в корректности внесённых им изменений в словари. При закрытии формы поиска предлагается сохранить изменения в файлах словарей.

В процессе поиска присутствие разделителей справа и слева от искомой последовательности букв не учитываются. На данном этапе такое расширение результатов поиска может оказаться полезным для исследования морфологии старорусского языка.

При этом выделяется подсветкой наиболее длинная из искомых словоформ в данном включении в текст. Поиск остальных, более коротких словоформ в данном фрагменте не производится – определённую роль в оптимизации данного аспекта алгоритма играет упорядоченность словаря псевдоокончаний по убыванию их длины.

2.4 Технические особенности

Важным аспектом для применения результатов данной работы является интеграция со стандартными программными средствами. Локальный поиск для старорусского языка реализован в качестве макросов устанавливаемой надстройки-расширения для *MS Word* (дополнительное меню, панель инструментов и другие элементы). Данная надстройка помимо поиска включает ряд других функций, необходимых для работы лингвиста со старорусскими текстами.

Морфологические словари загружаются при первом запуске формы поиска и содержатся в оперативной памяти до тех пор, пока не будут закрыты все окна экземпляра приложения *MS Word*. На «слабом» компьютере Intel Pentium III 800 EB (801 МГц) 256 кБ, DIMM 128 МБ PC133 (133 МГц) для словаря псевдооснов объёмом 143450 слов в кодировке Unicode (3 МБ) загрузка требует 3-х секунд. Время загрузки словаря окончаний несущественно.

Опционально поиск осуществляется: с учётом словоизменения, с учётом регистра, в выделённом фрагменте, в прямом или обратном направлении. Предусмотрено возвращение на исходную позицию. Кроме текущего найденного слова, выделенного в документе подсветкой, пользователь наблюдает в форме список всех найденных в документе слов, соответствующих запросу, в порядке их следования. Результаты поиска по документу для каждого запроса можно сохранить в отдельный файл. Допустимые размеры словарей псевдооснов и псевдоокончаний ограничены объёмом примерно 2×10^9 записей.

Меню надстройки для *MS Word* представляет возможность отправки отредактированных данных словаря псевдооснов и псевдоокончаний через сеть Интернет для согласования с разработчиками и использования в следующих версиях программы и пакетах обновления.

Код написан преимущественно на *VBA* с использованием *Win32API* и снабжён подробными комментариями. Имена переменных соответствуют венгерской нотации.

2.5 Эксперименты

2.5.1 Генерация потенциальных псевдооснов

Стемминг слов *SAP* для псевдоокончаний старорусского языка позволил выделить неполный набор потенциальных псевдооснов. Для каждого слова *SAP* было порождено в среднем 2,4 гипотетических псевдоосновы. При этом стоит

обратить внимание, что в САР большинство слов являются леммами, что снижает количество возможных псевдооснов.

Следует отметить, что для современного русского языка для одной не словарной словоформы парсер *MyStem* порождает в среднем 3 гипотезы морфологического разбора [29].

2.5.2 Тестирование функции поиска

Несмотря на определённую вариативность морфологий, алгоритм поиска корректно работает для текстов начала и конца XVIII века.

Тестирование проводилось в двух режимах: без учёта разделителей между словами текста и с учётом. Первый режим был необходим для широкого мониторинга работы программы, исследования состава текстов, словарей старорусского языка и статистики «омонимичного» включения псевдооснов. Второй режим был необходим для качественной оценки работы программы. Ниже описано тестирование во втором режиме.

Целью тестирования было определить, является ли предложенный алгоритм корректным для работы с текстами различных жанров для всего XVIII века. Тестирование проводилось с использованием словаря псевдооснов современного русского языка, поскольку использование генерированного словаря псевдооснов старорусского языка мало отличается от чисто эвристического подхода, а ручная коррекция словаря требует времени. В соответствии с алгоритмом, если основа не была определена, применялась эвристика.

В эксперименте использовалась генерация словарей, разработанная в качестве одной из функций надстройки *MS Word*: для каждого старорусского текста, который использовался для исследования, создан словарь словоупотреблений (экземпляров слов). Пользуясь данным словарём, имея базовые знания о словоизменении старорусского языка, для каждого слова-запроса был составлен частотный список словоформ, содержащихся в тексте. Данная ручная выборка производилась в рамках словоизменения, не затрагивая словообразование (учитывая, что разница между ними условна). Затем производилось тестирование программы, и результаты поиска сопоставлялись со списками словоформ.

Для каждого используемого в исследовании текста было отобрано определённое количество первых в тексте слов, представленных более чем одной словоформой. Эти слова использовались в качестве запросов. Эксперименты повторялись для каждого следующего слова до тех пор, пока статистика результатов исследования по тексту не достигала относительной устойчивости. Как правило, было достаточно, чтобы исследование затронуло около 2% слов текста.

Если псевдооснова не была найдена в словаре, эвристический поиск иногда давал в качестве результатов помимо словоизменительных форм

некоторые словообразовательные формы.

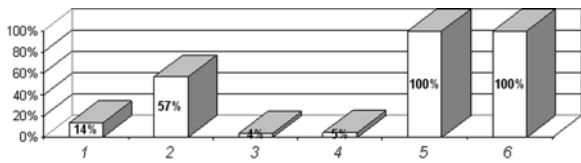
В процессе тестирования пополнение словаря псевдооснов и словаря псевдоокончаний не производилось, так как целью тестирования было также определить качество эвристического анализа и объём требуемых дальнейших работ по расширению словарей.

Для каждого исследованного текста составлена диаграмма результатов поиска. На всех представленных ниже диаграммах первый столбец показывает, для какого процента слов-запросов стемминг был произведён некорректно по причине недостаточной полноты словаря псевдоокончаний. Второй столбец показывает, для какого процента слов-запросов псевдооснова не была найдена в словаре псевдооснов современного русского языка. Третий столбец показывает, какой процент искомым словоформ, присутствующих в тексте, не был найден. Четвёртый столбец показывает, какой процент искомым словоупотреблений не был найден. Пятый столбец показывает, какой процент не найденных программой словоформ не был найден по причине чередования псевдооснов. Шестой столбец показывает тот же процент для словоупотреблений. Данное графическое представление, без группировки и наложения столбцов, оказалось наиболее наглядным для выделения существенных характеристик.

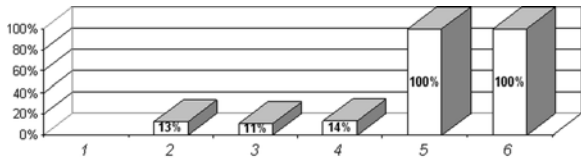
Таким образом, для двух первых столбцов процент соответствующих ошибок поиска вычисляется относительно слов-запросов, для 3-го и 4-го столбцов процент ошибок вычисляется относительно количества слов, которые должны были быть найдены в лучшем случае, для 5-го и 6-го столбцов процент ошибок вычисляется соответственно от общего количества не найденных словоформ и словоупотреблений. Указанные в диаграммах проценты округлены до целых. Для категорий, имеющих значение 0%, столбцы отсутствуют. Наиболее показательными являются 2-й, 3-й и 4-й столбцы.

Выбор текстов для исследования был ограничен в связи с отсутствием точной датировки многих произведений. Результаты работы с текстами 1-й трети XVIII века представлены на диаграммах 1-4. Тексты, для которых составлены данные диаграммы, написаны различными авторами и представляют собой церковные грамоты. Работа с самым ранним из имеющихся в распоряжении текстов XVIII века [4] (диаграмма 1), демонстрирует высокое качество эвристического анализа при малом количестве словарных псевдооснов. Диаграмма 2 отражает статистику одновременно по двум историческим документам [7,3], близким по содержанию, связанным с текущими историческими событиями и написанным в один день, но различными авторами. Статистика для данных произведений совпадает. В тот же день выходит ещё одна грамота [5], статистика по которой представлена в диаграмме 3, и в тот же месяц – ещё

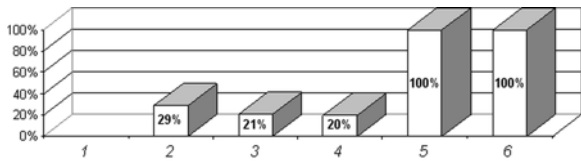
одна [6] (диаграмма 4). Определённое сходство данных текстов отражается в диаграммах.



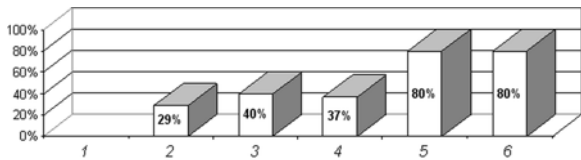
Диаг. 1. Грамота Петра I, 1721 г.



Диаг. 2. Грамоты Иеремии и Афанасия, 1723 г.

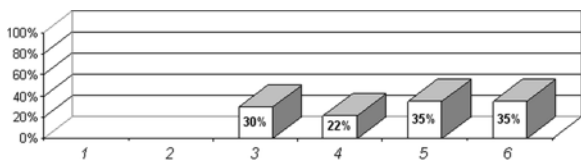


Диаг. 3. Грамота вселенского патриарха, 1723 г.

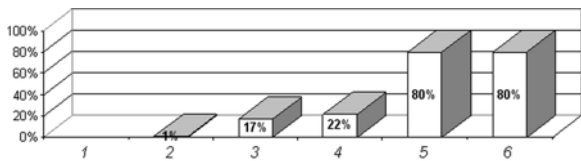


Диаг. 4. Грамота восточных патриархов, 1723 г.

Вторая треть XVIII века представлена в исследовании двумя письмами М.В.Ломоносова к И.И.Шувалову [15,16] с разницей в 8 лет (диаграммы 5 и 6 соответственно). Однако прямое сходство между данными диаграммами не наблюдается.

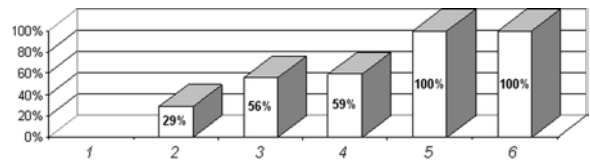


Диаг. 5. Письмо Ломоносова Шувалову, 1753 г.

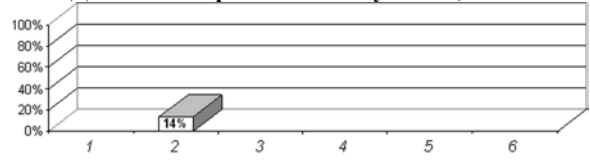


Диаг. 6. Письмо Ломоносова Шувалову, 1761 г.

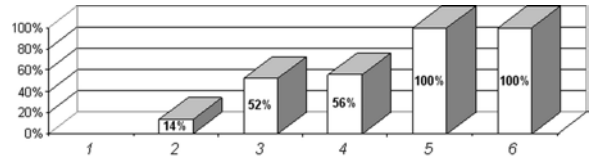
Конец XVIII века представлен на диаграммах 7-9 произведениями Крылова, Раstopчина и Батюшкова [1,13,26]. Произведения созданы с разницей во времени максимум 2 года и относятся к различным жанрам: поэзия, историческое повествование и эпистолярный жанр. Высокая релевантность поиска в [26] является исключением.



Диаг. 7. «Подражание псалму XVII», 1795 г.

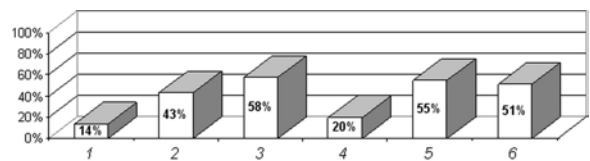


Диаг. 8. «Последний день Екатерины II», 1796 г.

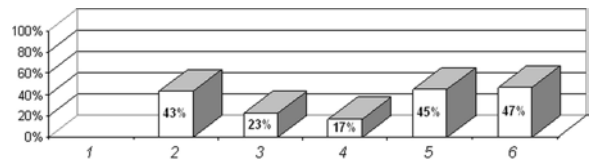


Диаг. 9. «Сётрамъ», 1797 г.

Кроме того, были проведены исследования для текстов, принадлежащих так называемому долгову XVIII века (1660-1825 г.) [34,35] – Диаграммы 10, 11. Понятие «долгий XVIII век» используется специалистами по истории русского языка. Для такого исследования в распоряжении имеется два произведения с точной датировкой перевода, принадлежащие жанру поэзии. Оба произведения отличаются сравнительно низкой долей ошибок, вызванных чередованием псевдооснов.

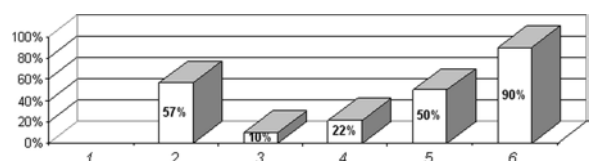


Диаг. 10. «Россияда», 1807 г.

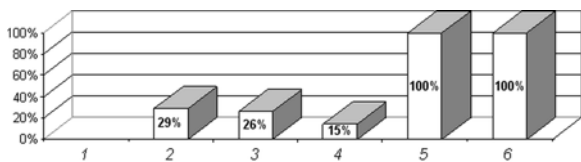


Диаг. 11. «Песнь российскому слову», 1809 г.

Отдельно требуется рассмотреть переводы с иностранных языков, выполненные в XVIII веке. В рамках проекта было оцифровано и переведено в текстовый формат (с использованием программы *ABBYY FineReader* и посткоррекции) историческое повествование [12] (диаграмма 12), – перевод с латинского языка. Для исследования было также использовано сатирическое произведение [27] (диаграмма 13) – перевод с греческого и латинского языков.



Диаг. 12. «Книга система...», 1722 г.



Диаг. 13. «Война мышей и лягушек», 1772 г.

Как видно из диаграмм, качество поиска с учётом словоизменения XVIII века на основе предложенного алгоритма не обладает большой зависимостью от времени создания произведения. Не смотря на то, что некоторое влияние темы, жанра и времени создания произведений прослеживается практически во всех случаях, основной общей причиной ошибок поиска является чередование псевдооснов.

Статистика диаграмм показывает, что проблема чередующихся основ является наиболее актуальной. Для повышения качества поиска необходимо индексировать словарь основ и определить парадигмы словоизменения, а также, возможно, использовать классифицирующие грамматические категории. В целях упрощения данной работы была создана исходная версия грамматической разметки САР. С использованием разработанной программы для каждого слова САР осуществлён поиск его современного варианта в электронной версии словаря А.А.Зализняка и перенос грамматических индексов. Таким образом, было корректно найдено соответствие современным словам более чем для 50% старорусских слов САР. Помимо пополнения словаря, в связи с тем, что определённый процент грамматических характеристик старорусского и современного языка отличается, созданный словарь планируется подвергнуть коррекции.

2.6 Дальнейшие планы исследования

Совместно с сотрудниками кафедры ИРЯЯ КГУ планируется осуществить пополнение и модификацию словарей. Базовый словарь планируется структурировать аналогично морфологическому словарю системы «Диалинг» и индексировать соответственно грамматике старорусского языка. Разметка словаря позволит усовершенствовать алгоритм поиска. Планируется создание программы для работы с морфологическим словарём старорусского языка в кодировке Unicode. Компоненты системы «Диалинг» для данных целей не подходят.

2.7 Перспективы внедрения

Внедрение представляется возможным в расширенном поиске *Яндекс*, в Национальном корпусе русского языка (НКРЯ) [20] на сервере *Яндекс*, а также в настольном приложении «Персональный поиск» при условии, что в нём будет в достаточной мере реализована поддержка Unicode.

Внедрение морфологического расширения для старорусского поиска в систему *Яндекс* представляется более универсальным решением,

чем разработка специальной метапоисковой системы на базе *Яндекс.XML*. Символы запроса на старорусском языке корректно обрабатываются из строки расширенного поиска *Яндкса*. Как и для других языков, можно автоматизировать выбор морфологического расширения старорусского языка, определяя его в запросе по характерным символам XVIII-XIX веков. При этом также необходимо отключение автоматического перехода в интерфейс стандартного поиска *Яндекс*: эта особенность принуждает возвращаться к расширенному поиску при каждом следующем запросе в старорусской орфографии. В простейшем случае представляется возможным ввести соответствующую опцию. В лучшем случае можно реализовать достаточно полную поддержку Unicode в стандартном поиске *Яндекс*.

Некоммерческая версия парсера *MyStem* некорректно работает с текстовыми файлами в кодировке Unicode. Однако *MyStem* правильно обрабатывает десятичные Unicode-коды XML (например, ѣ – «ять» и т.д.). Перед морфологическим анализом *MyStem* заменяет старорусские символы родственными символами современного языка («ять» – «е» и т.д.) и производит анализ соответственно словоизменению современного русского языка. В отличие от некоммерческой версии парсера *MyStem* расширенный поиск *Яндекс* для старорусского языка производится точно как в запросе.

3. Выводы и перспективы

Исследование показало, что поиск с учётом морфологии в старорусских текстах является алгоритмически и технически возможным и достаточно эффективным.

Более широкое тестирование созданных алгоритмов может быть проведено в рамках НКРЯ. С этой целью подготовлены тексты XVIII века в оригинальной орфографии для включения в состав НКРЯ.

Созданная программа, исходный код, словари и тестовые образцы текстов представлены на сайте [23] в свободном доступе. В дальнейшем планируется интеграция проекта с сайтом библиотеки КГУ [22], где в настоящее время также размещены некоторые разработки. Все разработанные ресурсы могут быть использованы для создания корпусов текстов XVIII века.

4. Благодарности

Исполнители данного проекта выражают благодарность заведующему каф. истории русского языка и языкознания КГУ Николаеву Г.А. за предоставленные консультации, а также зав. отд. редких книг и рукописей библиотеки КГУ им. Лобачевского Амерхановой Э.И. за предоставленные материалы и доступ к фондам.

5. Литература

- [1] Батюшков К.Н. Сёстрамъ / К.Н.Батюшков // Фундаментальная электронная библиотека "Русская литература и фольклор". – Режим доступа: <http://feb-web.ru/feb/batyush/texts/ps0/ps3/ps320012.htm>.
- [2] Богатова Г.А. Словарь Академии Российской 1789-1794. – Т.1-6 / гл. ред. Г.А.Богатова // М.: МГИ им Е.Р.Дашковой. – 2001-2006.
- [3] Грамата Аѳанасія, Патріарха Великаго Божія града, Антіохіи и всего Востока // Проект «Наследие святой Руси». – Режим доступа: <http://nasledie.russportal.ru/index.php?id=history.synod02>.
- [4] Грамата Императора Петра Перваго къ Патріарху Константинопольскому Іереміи, объ учрежденіи въ Россіи Свягѣйшаго Синода // Проект «Наследие святой Руси». – Режим доступа: <http://nasledie.russportal.ru/index.php?id=history.synod01>.
- [5] Грамата Іереміи, Архіепископа Константина града, новаго Рима, и Вселенскаго Патріарха // Проект «Наследие святой Руси». – Режим доступа: <http://nasledie.russportal.ru/index.php?id=history.synod02>.
- [6] Грамата Іереміи, Аѳанасія и Хрисанѳа // Проект «Наследие святой Руси». – Режим доступа: <http://nasledie.russportal.ru/index.php?id=history.synod02>.
- [7] Грамата Іереміи, Патріарха Константина града // Проект «Наследие святой Руси». – Режим доступа: <http://nasledie.russportal.ru/index.php?id=history.synod02>.
- [8] Издательский дом «Русскій Паломникъ». Официальный сайт [Электрон. ресурс]. – Режим доступа: <http://idgrp.ru>.
- [9] Интернет-сообщество «Восемнадцатый век». [Электрон. ресурс]. – Режим доступа: http://community.livejournal.com/18century_ru.
- [10] Интернет-сообщество «Ижица». [Электрон. ресурс]. – Режим доступа: http://community.livejournal.com/ijitsa_ru.
- [11] Ирмологий [Проект по разработке шрифтов Электрон. ресурс]. – Режим доступа: <http://irmologion.ru>.
- [12] Кантемир Д.К. Книга свѣста или состояніе мухаммеданскія религіи / Д.К.Кантемир // Түпографія царствующаго Санктъ-Петербурха. – С-Пб. – 1722. – 412 с.
- [13] Крыловъ И.А. Подражаніе псалму XVII / И.А.Крыловъ // «Im Werden Verlag». Некоммерческое электронное издание. – Режим доступа: http://imwerden.de/pdf/krylov_duhovnye_stixotvoreniya.pdf.
- [14] Лингвистический Интернет-форум [Электрон. ресурс]. – Режим доступа: <http://lingvoforum.net>.
- [15] Ломоносов М.В. Письмо М.В.Ломоносова къ И.И.Шувалову 01.11.1761 / М.В.Ломоносов // «Im Werden Verlag». Некоммерческое электронное издание. – Режим доступа: http://imwerden.de/pdf/lomonosov_pismo_shuvalovu_1.11.1761.pdf.
- [16] Ломоносов М.В. Письмо М.В.Ломоносова къ И.И.Шувалову 1753 г. / М.В.Ломоносов // «Im Werden Verlag». Некоммерческое электронное издание. – Режим доступа: http://imwerden.de/pdf/lomonosov_shuvalovu_1753.pdf.
- [17] Манускрипт. Информационно-поисковая система [Электрон. ресурс] // Лаборатория по автоматизации филологических работ УдГУ. – Режим доступа: <http://manuscripts.ru>.
- [18] Маргулис И.С. Кодовое пространство и лексикографическая сортировка для языков, использующих славянскую графику / И.С.Маргулис // Исследования по информатике. – Казань: Отечество, 2007. – №11. – С.129-152.
- [19] Маргулис И.С. Раскладка клавиатуры для работы с русскоязычными текстами XVIII-XIX веков / И.С.Маргулис // Исследования по информатике. – Казань: Отечество. – 2005. – Вып.9. – С.133-138.
- [20] Национальный корпус русского языка // Ассоциация «Национальный корпус русского языка» [Электрон. ресурс]. – Режим доступа: <http://www.ruscorpora.ru>.
- [21] Официальный сайт Института русской литературы РАН (Пушкинский дом) [Электрон. ресурс]. – Режим доступа: <http://pushkinskiydom.ru>.
- [22] Официальный сайт Научной библиотеки им. Н.И.Лобачевского Казанского государственного университета им. В.И.Ленина. [Электрон. ресурс]. – Режим доступа: <http://isl.ksu.ru>.
- [23] Проект «Автоматическая обработка старорусских текстов» [Электрон. ресурс]. – Режим доступа: <http://18.ucoz.ru>.
- [24] Проект «Автоматическая обработка текста» [Электрон. ресурс]. – Режим доступа: <http://www.aot.ru>.
- [25] Проект «Русскій поискъ» [Электрон. ресурс]. – Режим доступа: <http://www.poisk.russian.ru>.
- [26] Растопчинъ Ѳ.А. Послѣдній день жизни императрицы Екатерины II и первый день царствованія Императора Павла I / Ѳ.А.Растопчинъ // Проект «Наследие святой Руси». – Режим доступа: <http://nasledie.russportal.ru/index.php?id=histrus.rastopchin>.
- [27] Рубан В.Г. Война мышей и лягушекъ / В.Г.Рубан. // «Im Werden Verlag». Некоммерческое электронное издание [Электрон. ресурс]. – Режим доступа: http://imwerden.de/pdf/ruban_vojna_myshej_i_lyagushek.pdf.
- [28] Русский филологический портал [Электрон. ресурс]. – Режим доступа: <http://www.philology.ru>.
- [29] Сегалович И. Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов / И.Сегалович, М.Маслов // Диалог'98. Казань, 1998. Т.2. С. 547-552.
- [30] Словарь Академіи Россійской. – Т.1-6. – С-Пб.: Императорская Академія Наукъ, 1789-1794.
- [31] Сообщество «Письменное наследие». Официальный сайт [Электрон. ресурс]. – Режим доступа: <http://textualheritage.org>.
- [32] Старостин С.А. Проект «Эволюция языка» [Электрон. ресурс] / С.А.Старостин [и др.]. – Режим доступа: <http://starling.rinet.ru>.

- [33] Филиппович А.Ю. Информационная технология создания электронного издания Словаря Академии Российской 1789-1794 гг. / А.Ю.Филиппович // *Материалы Междунар. научн. конф. «Современные информационные технологии и письменное наследие»* (Ижевск, 13-17 июля 2006). – Ижевск: Изд-во ИжГТУ. – 2006. – С.174-178.
- [34] Херасковъ М.М.. Россияда / М.М.Херасковъ // «Im Werden Verlag». Некоммерческое электронное издание. – Режим доступа: http://imwerden.de/pdf/kheraskov_rossiada.pdf.
- [35] Шихматовъ С.А. Пѣснь российскому слову / С.А.Шихматовъ // «Im Werden Verlag». Некоммерческое электронное издание. – Режим доступа: http://imwerden.de/pdf/shikhmatov_pesn_rossijskomu_slovu.pdf.
- [36] Энциклопедия «Википедия» [Электрон. ресурс]. – Режим доступа: <http://ru.wikipedia.org>.
- [37] Энциклопедия культур. Авторский проект Александра Бокшицкого [Электрон. ресурс]. – Режим доступа: <http://ec-dejavu.ru>.
- [38] Южиков В.С. Об одном методе предварительной обработки изображений старопечатных и рукописных текстов / В.С.Южиков // *Исследования по информатике*. – Казань: Отечество. – 2005. – Вып.9. – С.125-132.
- [39] *Philologica*. Двухязычный журнал по русской и теоретической филологии [Электрон. ресурс]. – Режим доступа: <http://www.rvb.ru/philologica>.
- [40] Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine / I.Segalovich // Публикации сотрудников ООО «Яндекс» [Электрон. ресурс]. – Режим доступа: <http://company.yandex.ru/articles/iseg-las-vegas.html>.
- [41] Solovyev V. Electronic Library of Russian Books of the XVIII Century: Problems and Perspectives / V. Solovyev // *Proc. International conf. London-EVA'05*. London. 2005. p.329-335.

XVIII century were compiled during the project implementation for the first time. We developed the program for local search of words of ancient spelling taking into account word-changes for the Russian language of the XVIII century. The program is an installed extension for MS Word. We have been working on text corpuses of the XVIII century, in which the results of the work may be used. The research is supported by RFBR and RFH.

Search in Russian Text Collections of the XVIII Century

Solovyev V., Margulis I.

The investigation of Old-Russian texts becomes much easier with the use of their digital versions as the materials are acceptable not only for researches from big university centers. Digital versions of old texts are created by the biggest libraries of the world including the Russian State Library. The necessity and different aspects of digitization of old writing were analyzed in detail in the work *Solovyev V. Electronic Library of Russian Books of the XVIII Century: Problems and Perspectives. Proc. International conf. London-EVA'05. London. 2005.*

First of all the investigation of Old-Russian texts requires the development of morphological analyzers and search programs, oriented to word-changing in the Old Russian language. It looks much more promising to adapt contemporary algorithms of morphological analysis and search for the Old Russian language, using electronic dictionary of the XVIII century for solving the problem.

The direct and reversed electronic dictionaries of the