МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

Писменото наследство и информационните технологии [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.

ISBN 978–954–9787–25–2

Из опыта разработки автоматического морфологического анализатора для текстов XVIII–XIX века[1]

С. О. Савчук

Автоматический анализ текстов, грамматический словарь, аннотация диахронических корпусов, тексты XVIII–XIX века

The Experience of Developing an Automatic Morphological Analyzer for 18th– to 19th-century Texts

Svetlana Savchuk

This paper presents the key principles of building a morphological analyzer for 18th– to 19th-century Russian texts. The analyzer should involve different modules applicable to different kinds of texts depending on their respective orthographical and grammatical phenomena. Evaluation data of the results of the first analysis are presented.

Создание морфологического анализатора для текстов исторической части Национального корпуса русского языка является в настоящий момент актуальной задачей, поскольку тексты, относящиеся к XVIII, XIX и 1-й пол. XX вв., составляют более 110 млн словоупотреблений. Автоматический анализ на основе современного словаря дает неплохие результаты для текстов вплоть до 2-й пол. XVIII в. (Ломоносов, Карамзин и др.). Неопознанные анализатором слова с пометой *bastard* мы собираем в частотные списки словоформ, заносим в специальную базу вариантов, наиболее частотные формы с приписанными грамматическими признаками включаются в специальные конфигурационные файлы add.cfg, и при индексации текстов эти словоформы в текстах также получают правильные разборы. Так, в частности, были решены проблемы с формами на *-ию* (*радостию*, *робостию*), *-ся* (*улыбалося*, *казалося*), деепричастиями (*идучи*, *вышед*, *познакомясь*) и нек. др.

Однако настал момент, когда такие полумеры уже не спасают положение. Происходит это по нескольким причинам. Во-первых, меняется состав исторической части корпуса: в него все больше включаются тексты с повышенным содержанием нестандартных словоформ — тексты начала XVIII в., дневники и письма с ненормированной орфографией. Во-вторых, активно создается подкорпус текстов в оригинальной орфографии, который формируется из отсканированных дореволюционных изданий. Наконец, в-третьих, старорусский корпус,

[1] Работа выполнена при поддержке: Программы фундаментальных исследований Президиума РАН “Корпусная лингвистика”.

объемом около 4 млн словоупотреблений, нуждается в морфологической аннотации, которая на таком массиве может быть осуществлена только автоматически. Таким образом, необходимость создания морфологического анализатора для исторической части корпуса стала очевидной. Анализатор должен удовлетворять следующим общим требованиям: иметь гибкие настройки на различные классы текстов, давать приемлемые результаты для текстов ранних периодов, иметь возможность использования в полностью автоматическом и полуавтоматическом режимах.

В настоящее время создана рабочая версия анализатора, которая была протестирована на экспериментальном корпусе текстов XVIII–XIX вв. Результаты тестирования описаны в статье [Поляков и др. 2013] и частично внедрены в новую версию анализатора.

Что представляет собой программа-анализатор? В программе можно выделить два блока — словарь и парсер. Словарь состоит из двух частей — словаря флексий с приписанными граммемами и словаря основ с указанными чередованиями и парадигмами словоизменения. Словник словаря пока совпадает со словником грамматического словаря А. А. Зализняка, но в дальнейшем предполагается его расширение.

Программа-парсер опознает словоформу в тексте, вычленяет в ней основу и флексию, приписывает правильную лемму и грамматические признаки из словаря. Относительно отсутствующих в словаре форм программа строит гипотезы об их лемме и грамматических признаках, при этом существует возможность ограничивать максимальное количество порождаемых гипотез. Поскольку программе придется анализировать тексты в дореволюционной орфографии, в ней учтены графические и орфографические правила, устанавливающие соответствия между реальным, представленным в тексте написанием и современным нормализованным. Что умеет анализатор?

Анализатор поддерживает орфографию XVIII–XIX вв. (oldspell) и автоматически преобразует дореформенную орфографию в современную. Перечислим некоторые правила, реализованные в настоящей версии.

1. ять => е, фита => ф, i/v => и.

2. -ъ => пусто.

3. без-/в(о)з-/из-/низ-/раз-/роз-/ч(е)рез- => *-с* перед глухими.

4. Парсер заменяет некоторые недопустимые сочетания букв на правильные:

– ъи => ы, ъ + [аоэу] => пусто (*съиграть*, *съузить*, *съэкономить*);

– [чшщж] + [яюы] => [ауи] (*чящю*, *жыву*).

5. Некоторые частотные орфографические варианты включены в парадигмы (с пометой *orfo*):

– для текстов периода XVIII в. (old2): ч+ь (врачь); +еш, +иш (*нес+еш*, *ход+иш*);

– для периода XIX в. (old1): +о/е после [чшщжц] (*отц+ем*, *отц+ев*, *пальц+ом*, *пальц+ов*, *душ+ею*) и др.

Анализатор учитывает словоизменение XVIII–XIX вв. (oldrus). В настоящей версии внедрены следующие правила.

Для всех периодов (режим parser.options.oldrus = 0)

1. Вариант частицы *-ся* после гласных (*валю+ся*, *валила+ся*), который употребляется в современном языке.
2. Деепричастия совершенного вида от основы презенса (*прийдя*, *увидя*, *взгромоздясь*), которые вполне употребительны в современном языке, но не учтены в словаре Зализняка.
3. Сравнительная степень на *-ѣй* (*сильн+ѣй*).

Для периода XIX в. (режим parser.options.oldrus = 1)

1. Адъективные флексии (*+аго/яго*, *+ыя/ия*).
2. Творительный падеж 3-го склонения на *-ию* (*милост+ию*, *помощ+ию*).
3. Мест. ед. среднего рода на *-и* (*о копь+и, варень+и, здань+и*).
4. Особые формы местоимений (*ея*, *он+ѣ*, *одн+ѣ*, *одн+ѣхъ*).

Для периода XVIII в. (режим parser.options.oldrus = 2)

1. Усеченные формы прилагательных (*красн+а/о/ы/у*).
2. Сравнительная степень на *-яе/ае* (*сильн+яе, чуж+ае*).
3. Глагольные флексии *-ти* и *-ши* (*ходи+ши*, *ходи+ти*).
4. Множ. число среднего рода на *-ы/-и* (*озер+ы*, *войск+и*, *лиц+ы*).

Для более ранних периодов и церковнославянских текстов (режим parser.options.oldrus = 3)

1. Формы множественного числа существительных (*град+и*, *град+ом*, *град+ѣх*, *град+ми*, *кон+ьми*).
2. Флексия *-а* вместо *-я* после *-и* (*Ефреми+а*, *здани+а*).
3. Формы аориста (*ходи+х*, *ходи*, *ходи+хом/сте/ша*, *пек+ох*, *печ+е*, *пек+охом*, *пек+осте*, *пек+оша*) и имперфекта (*любл+ях*, *любл+яше*, *любл+яхом*, *любл+ясте*, *любл+яху*, *печ+ах*, *печ+аше*, *печ+аху*) и др.

Текущая версия программы была протестирована на текстах, относящихся к разным периодам и типам (жанрам), в оригинальной и современной орфографии. Для слов, отсутствующих в словаре анализатора и получающих гипотетические разборы, было установлено ограничение в 10 гипотез. Словоформа считалась опознанной правильно, если приписанная лемма и одна из 10 предложенных гипотез оказывалась правильной. Анализ результатов тестирования приводится ниже.

Для текстов XIX в. в оригинальной орфографии (несколько писем А. С. Пушкина по изданию Модзалевского, тексты из журналов “Современник” и “Москвитянин” за 1850 год) результаты очень хорошие. Программа справилась с оригинальной дореволюционной орфографией: до 99 % словоформ получили правильные разборы. Неправильные разборы в письмах Пушкина получили ор-

фографические “ошибки” — слитное *нехочется*, *понадобиться* вместо *понадобится*, *за чем* вместо *зачем*. В журнале “Современник” и “Москвитянин” ошибочные разборы получили словоформы *(с) краю* (дат. п. вместо род. п.), *сумерокъ* (род. п. существительного *сумерки* вм. нормативного *сумерек*), *блещутъ* (неправильно определена лемма).

Для текстов XVIII в. картина примерно аналогичная. Тексты из журнала “Лекарство от скуки и забот” за 1798 год в оригинальной орфографии дали высокий процент правильно разобранных словоформ — 98 %, ошибки отмечены в определении леммы и морфологических характеристик форм *еслибъ*, *похочет* и имени *Иоанъ*. Текст Радищева (в модернизированной орфографии), стиль которого отличается обилием архаизмов и книжных слов, тем не менее показал 98 % правильных разборов. Не опознаны *егда*, *еже*, *идеже*, которые войдут в расширенный словарь, форма *вихрящася* (деепричастие от глагола *вихриться*).

Отдельной задачей было оценить возможности использования анализатора для разметки старорусского корпуса. Результаты точечной проверки нескольких текстов XVII в. показали следующее. Фрагмент текста “Путешествие стольника П. А. Толстого по Европе. 1697–1699” (по изданию 1992 г.) дал неожиданно высокие результаты — 97 % правильно разобранных форм и гипотез. Большая часть ошибочных разборов — орфографические варианты написаний форм *ево*, *никакова*, *приезжей* (им. ед. муж. р. вм. *приезжий*), *толко*, *водак* (род. мн. от *водка* вм. *водок*), *Венеци*, *Франци* вместо *Венеции*, *Франции* и нек. др. Кроме того, шире представлены морфологические варианты слов, с которыми автоматически справиться не удастся, например *ужина*, *ужиною*, *ужины* свидетельствуют о существовании в XVII–XVIII в. варианта жен. рода к современному слову *ужин*.

Анализ образца деловой письменности — “Царская грамота кунгурскому воеводе Алексею Калитину” (Кунгурские акты XVII века (1668–1699 гг.). СПб., 1888) — показал 95 % правильных разборов и гипотез. Но на этот раз преобладали трудности грамматического порядка: устаревшие формы *по сту рублевъ*, *посаду*, *уезду* (род. п.), *делы* (твор. мн.), *Тотемцомъ* (дат. мн.) не были проанализированы правильно, потому что отсутствовали в парадигмах. В деловых жанрах сразу можно предсказать трудности иного порядка, обусловленные обилием собственных имен — личных имен и топонимов. Если список личных имен еще както ограничен, то в совокупности с фамилиями, отчествами, прозвищами он становится огромным.

Как представляется, результаты небезнадежные. Несомненно, пополнение словника материалом старых словарей и расширение списка правил для раннего периода повысит качество результатов. Но уже сейчас ясно, что анализатор можно использовать с целью предварительной автоматической разметки текстов среднерусского корпуса, за которой последует ручное постредактирование. В процессе такого ручного анализа будут накапливаться решения, которые могут быть внедрены в анализатор и будут служить совершенствованию качества разметки.

Литература

Поляков и др. 2013 — *Поляков А. Е., Савчук С. О., Сичинава Д. В.* Грамматический словарь для автоматического анализа текстов XVIII–XIX века: первые результаты // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции “Диалог” (Бекасово, 29 мая – 2 июня 2013 г.). Вып. 12 (19). М.: Изд-во РГГУ, 2013. С. 632–654.