Вестник ПСТГУ. Серия III: Филология
2016. Вып. 2 (47). С. 7–25

Гаврилова Татьяна Сергеевна,
НИУ «Высшая школа экономики»
tanya96gavrilova@yandex.ru@mail.ru

Шалганова Татьяна Александровна,
НИУ «Высшая школа экономики»
insana7@mail.ru

Ляшевская Ольга Николаевна,
канд. филол. наук,
НИУ«Высшая школа экономики»,
Институт русского языка им. В. В. Виноградова РАН
olesar@yandex.ru

К задаче автоматической лексико-грамматической разметки старорусского корпуса XV–XVII вв.[*]

Т. С. Гаврилова, Т. А. Шалганова, О. Н. Ляшевская

В работе рассматриваются два подхода к разработке автоматической аннотации корпуса старорусских текстов XV–XVII вв., включенных в Национальный корпус русского языка (НКРЯ). Лексико-грамматическая аннотация состоит в определении части речи, грамматических характеристик и начальной формы слова (леммы) для каждой словоформы корпуса. Язык старорусской письменности совмещает в себе, с одной стороны, черты древнерусского словоизменения, включая формы аориста и имперфекта глагола, формы двойственного числа и другие архаичные формы, а с другой стороны — черты современной русской словоизменительной морфологии. Подобное смешение проявляется и в лексическом составе. Кроме того, в текстах присутствуют церковнославянские, а также диалектные варианты. Если добавить к этому отсутствие устойчивой орфографии, становится понятна вся сложность задачи, связанной с аннотацией старорусских текстов.

Первый из рассматриваемых подходов основан на построении электронного словаря старорусского языка и создании модуля обработки орфографической вариативности. В отсутствие открытых электронных ресурсов, документирующих морфологию старорусского периода, за основу был взят электронный словарь церковнославянского языка, разработанный А. Е. Поляковым на базе церковнославянского корпуса НКРЯ. Мы описываем процедуры, связанные с адаптацией именной и глагольной морфологии к данным старорусского корпуса.

Второй подход связан с привлечением программы автоматической аннотации текстов русского языка XIX–XX вв., дополненной модулем обработки орфографической вариативности, с одной стороны, и корпуса лексико-грамматических разборов древнерусских текстов, полученных из Исторического корпуса НКРЯ, — с другой.

[*] Исследование выполнено при частичной финансовой поддержке РГНФ, грант № 15-04-12050 «Развитие Исторических модулей НКРЯ».

Оба подхода строятся на принципе «широкого покрытия»: автоматический разметчик должен порождать множество разборов таким образом, чтобы хотя бы один разбор был правильным.

В статье приводятся результаты экспертизы качества разметки, основанной на указанных подходах, а также обсуждаются возможные пути развития инструментов лексико-грамматической разметки старорусских текстов.

1. Об особенностях автоматической морфологической аннотации старорусских текстов

Несмотря на обилие качественных и достаточно точных инструментов обработки современных текстов, предпринимаются пока что еще первые робкие попытки в области разработки автоматических анализаторов для древних и «старых» языков (например, для латинского, древнегреческого, древнеанглийского, древневерхненемецкого языка и др.). Вместе с тем во всем мире набирает ход дигитализация рукописей и других документов, хранящихся в библиотечных и музейных фондах, в исторических архивах и коллекциях. Очевидно, что снабжение электронных версий текстовых документов корпусной разметкой открывает путь к созданию мощной информационной платформы для обеспечения исследований в области диахронии языка.

Старорусский корпус НКРЯ (http://ruscorpora.ru/search-mid_rus.html) включает тексты, написанные преимущественно с XV по XVII в., в том числе летописи и сказания, деловые документы, бытовую переписку, памятники религиозной литературы и др. В отличие от других исторических корпусов НКРЯ: древнерусского корпуса текстов XI–XIV вв.[1], корпуса берестяных грамот XI–XV вв.[2] и корпуса церковнославянских текстов XVII–XX вв.[3], — старорусский корпус в насто-

[1] В настоящее время древнерусский корпус насчитывает 443 тыс. словоупотреблений (14 текстов), интерфейс поиска: http://ruscorpora.ru/search-old_rus.html. Подробнее о корпусе см.: *Мишина Е. И., Пичхадзе А. А.* Древнерусский подкорпус Национального корпуса русского языка // Труды Института русского языка РАН им. В. В. Виноградова. Вып. 6, 2015. С. 99–115; *Молдован А. М.* Памятники древнерусской письменности в Национальном корпусе русского языка // Труды Института русского языка РАН. Вып. 6, 2015. С. 88–98; *Пичхадзе А. А.* Корпус древнерусских переводов XI–XII вв. и изучение переводной книжности Древней Руси // Национальный корпус русского языка: 2003–2005. М., 2005. С. 251–262; *Зобнин А. И., Пичхадзе А. А.* Корпус древнерусских переводов XI–XII вв.: результаты и перспективы // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2005. № 3. С. 44–47.

[2] Объем корпуса берестяных грамот 19,5 тыс. словоупотреблений (885 текстов). URL: http://ruscorpora.ru/search-birchbark.html. См.: *Сичинава Д. В.* Исторические корпуса Национального корпуса русского языка как инструмент диахронических исследований грамматики // Баранов В. А., Желязкова В., Лаврентьев А. М. (отв. ред.). Писменото наследство и информационните технологии: Материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.). София; Ижевск, 2014.

[3] Объем церковнославянского корпуса 4,7 млн. словоупотреблений (1254 текста). URL: http://ruscorpora.ru/search-orthlib.html. Подробнее см.: *Добрушина Е. Р., Поляков А. Е.* Корпус церковнославянского языка: возможности, методы создания, перспективы // Вестник ПСТГУ. Серия III: Филология. 2013. Вып. 1 (31). С. 32–44; *Поляков А. Е.* Корпус церковнославянских текстов в составе Национального корпуса русского языка, первая версия: проблемы и решения // Доклад на международной научной конференции «Информационные технологии

ящее время не имеет лексико-грамматической разметки. В нашу задачу входило разработать инструменты для автоматического определения словарной формы слова, его части речи и грамматических характеристик — таких, как род, число, падеж, время, степень сравнения и т. п. (без разрешения омонимии форм).

Хорошо известен ряд проблем, затрудняющих автоматический морфологический анализ текстов старорусского периода. Во-первых, это переходное состояние языка, совмещающего в себе разные грамматические и лексические слои[4]. Многие тексты отражают черты языка предшествующего периода (XI–XIV вв.), ср. такие формы, как *ꙗзыцѣ* (лемма *язык*, сущ. в местн. падеже ед. числа), *крилома* (лемма *крило*, сущ. в твор. падеже двойств. числа), *чюждїѧ* (лемма *чуждии*, прил. в род. падеже жен. рода ед. числа), *глагола* (лемма *глаголати*, глагол во 2–3-м лице ед. числа аориста) и др. Тексты также включают лексические или грамматические элементы церковнославянского языка. Нельзя упускать из вида и диалектное разнообразие текстов, написанных в разных диалектных зонах, ср., например, волог. *бусырь* ‘рабочая одежда’, сев. и сибирск. *исподка* ‘нижняя женская рубашка’. Все это означает, что для успешного анализа старорусских текстов компьютерная программа должна быть основана на «гибридной» морфологии, в частности она должна учитывать не только древнерусские и современные русские, но и церковнославянские лексемы и парадигмы.

Во-вторых, сложность представляет нестабильная орфография текстов старорусского периода[5] . В текстах корпуса наблюдается значительная вариативность письма, такая, например, как смешение *ѣ* и *е*, *ь* и *и, ѧ, я* и *ꙗ*; ср. написание формы прош. времени муж. рода глагола *взять* как *взꙗл, взѧлъ, възял, взялъ, взѧл, възялъ, взꙗлъ, взял, възѧлъ, вьзял, вьзял*; написание падежных форм существительного *польза* как *ползꙋ*, *ползоу*, *пользꙋ*, *пльза*, *по̀лзы*, *ползъ*; написание форм прилагательного *градьскыи* как *градцскаго* и *градстем*. Подобная вариативность может быть вызвана различными факторами, начиная с особенностей текста-первоисточника или диалектных особенностей области, в которой был создан текст, и заканчивая происхождением преподавателя, обучавшего носителя грамоте. Увеличивают орфографическое разнообразие сокращения под титлом и без него, а также варианты с выносными буквами (ср. *прпдбна*^{с} — лемма *прѣподобьныи, с̀крывша*^{с} — ма *съкрыти*). Свой вклад в графический разнобой вносит и то обстоятельство, что ряд текстов в корпусе воспроизводится по из-

и письменное наследие (El’Manuscript-12)». Петрозаводск, 2012; *Поляков А. Е.* Корпус церковнославянских текстов: проблемы орфографии и графики // Przegląd wschodnioeuropejski V/1, 2014, 245-254; *Добрушина Е. Р., Кравецкий А. Г., Поляков А. Е.* Корпус и частотный грамматический корпусный словарь церковнославянского языка в составе НКРЯ // Труды Института русского языка им. В. В. Виноградова. 2015. Вып. 6. С. 116–141.

[4] См.: Историческая грамматика русского языка: морфология; глагол / [Р. И. Аванесов, В. В. Иванов, В. Б. Силина и др.] Р. И. Аванесов, В. В. Иванов, ред. М.: Наука. 1982.; *Успенский Б. А.* История русского литературного языка (IX–XVII вв.). М.: Языки славянской культуры, 2002; *Живов В. М.* Очерки исторической морфологии русского языка XVII–XVIII веков. М., 2004; Историческая грамматика древнерусского языка /*В. Б. Крысько,* ред. Т. 1–4. М., 2000–2006, и мн. др.

[5] Ср., например, показательное исследование одного источника XVII в. в: *Демьянов В. Г.* Вести-Куранты: 1. Издание для исследования. 2. Исследование для издания //Лингвистическое источниковедение и история русского языка. М., 2000. С. 213–232.

даниям, в которых орфография рукописей упрощена (причем по разным правилам). Чем больше вариантов написания будет учтено при разработке ресурсов для аннотации, тем более полно и точно будут проанализированы тексты.

В-третьих, имеется техническая трудность — отсутствие ресурсов для автоматической разметки текстов старорусского периода. Анализаторы, созданные для обработки текстов современного русского языка, неприменимы к старорусским текстам отчасти из-за проблем, указанных выше (неоднородность морфологии и орфографическая вариативность), а отчасти из-за диахронических изменений, которые претерпела морфология языка. Из-за перестройки словоизменительных парадигм использование данных грамматического словаря А. А. Зализняка[6] будет заведомо недостаточным, а аналога такого словаря для более раннего периода развития русского языка, к сожалению, не существует.

Далее речь пойдет о разработке двух типов морфологических анализаторов (таггеров) для старорусского корпуса. Таггер представляет собой программу, способную проводить автоматическую лексико-грамматическую аннотацию текста, написанного на том или ином естественном языке. При помощи программы каждому грамматически корректному слову в тексте приписываются его начальная форма (лемма), часть речи и грамматические показатели (например, падеж, число и род у существительных, время, лицо, число и наклонение у глаголов и т. п.). В этой статье мы остановимся подробно на лексико-грамматических ресурсах, к которым может обращаться программа, а именно на новом грамматическом словаре текстов старорусского периода и на ранее размеченных вручную корпусных данных. Специально отметим, что другому большому блоку тем, а именно поддержке орфографических вариантов, от которой также зависит точность и объем покрытия разборов, посвящена отдельная наша работа[7].

Данная статья выстроена следующим образом. Во втором разделе будет представлен корпус, на основе текстов которого осуществляется разметка. В разделах 3–4 описывается словарный подход к разработке анализатора: третий раздел дает представление о принципах автоматического морфологического анализа и структуре представления данных в разрабатываемом электронном словаре, а четвертый раздел посвящен разработке именного и глагольного разделов электронного словаря. Пятый раздел представляет альтернативный подход, основанный на гибридном использовании существующих ресурсов (разборах древнерусского корпуса и грамматическом словаре современного русского языка). В шестом разделе мы приводим результаты оценки качества работы анализаторов на тестовом корпусе и подводим итоги.

---

[6] См.: *Зализняк А. А.* Грамматический словарь русского языка: Словоизменение. М.: Русский язык, 1977. 4-е изд., испр. и доп. — М.: Русские словари, 2003.

[7] См.: *Гаврилова Т. С., Шалганова Т. А., Ляшевская О. Н. Взіалъ, възялъ, вьзял* и т. д.: Обработка орфографической вариативности при лексико-грамматической аннотации старорусского корпуса XV–XVII вв. Рукопись. Москва, 2016. См. также: *Гаврилова Т. С.* Обработка вариативности при морфологическом анализе древнерусских текстов. Курсовая работа. Москва, НИУ ВШЭ, 2015; *Шалганова Т. А.* Создание грамматического словаря для морфологического анализатора древнерусского языка. Курсовая работа. Москва, НИУ ВШЭ, 2015.

2. Состав старорусского корпуса

Представленные в работе системы морфологического анализа ориентированы на рукописи, написанные в основном в период с XV по XVII в. Этот временной промежуток был выбран потому, что тексты древнерусского корпуса, созданные до XIV в., размечаются вручную, а тексты XVIII в. и более поздние могут быть обработаны при помощи инструментов, разработанных для современных русских письменных текстов (возможно, с некоторыми модификациями для анализа текстов XVIII в.[8]). Всего старорусский корпус содержит около 5000 текстов, однако большинство из них малы по объему. Наиболее широко в корпусе представлены официально-деловые тексты: грамоты, выписки из книг, записи, закладные кабалы, перечни, отписи, купчие, челобитные, данные, отписки, договорные грамоты, степенные книги и др. Деловую документацию дополняют тексты малых бытовых жанров: письма и грамотки. Значительный подкорпус составляют тексты религиозного содержания: жития, Четьи Минеи, поучения, песнопения, видения и т. п. Этих текстов не так много, как текстов делового содержания, однако они, как правило, велики по объему. К гибридным текстам в классификации жанров корпуса относятся летописи, хождения, поучения, наставления, заговоры и др. Подобное разнообразие жанров увеличивает сложности, возникающие при морфологическом анализе текстов. Тексты написаны в разных областях России, около трети текстов представляют собой списки с ранее созданных источников. Основная работа проводилась на корпусе объемом в 2 009 883 словоупотреблений. Далее в статье используется орфография текстов так, как она представлена в корпусе, со всеми разнообразными условностями и упрощениями, предпринятыми при публикации рукописей.

3. Принципы работы морфологического разметчика

Для разметки исторических корпусов языков флективного типа применяются как словарные системы разметки, так и статистические разметчики. Системы первого типа разрабатываются с использованием электронного словаря, в котором указаны основа (основы) лексемы и тип словоизменения, а также базы грамматических парадигм, в которой каждой грамматической форме, определяемой набором грамматических признаков, сопоставлено окончание, соответствующее тому или иному типу словоизменения. Системы второго типа, статистические, создаются на базе ранее размеченных корпусов с использованием скрытых марковских моделей, деревьев решений, машинного обучения на основе SVM и т. п.

Словарно-ориентированные анализаторы успешно зарекомендовали себя при разметке текстов современного русского языка: в качестве примера можно привести системы Mystem, АОТ, ЕТАР-3 и др., основанные на грамматическом словаре А. А. Зализняка. Примером применения словарных систем в разметке исторических текстов, написанных на индоевропейских языках, может служить

[8] См.: *Поляков А. Е.* Проблемы и методы анализа русских текстов в дореформенной орфографии // Компьютерная лингвистика и интеллектуальные технологии: По материалам Международной конференции «Диалог 2012». Вып. 11 (18). М., 2012. С. 536–547.

анализ существительных из корпуса древнечешского языка (Jínová et al. 2014). Система использовала грамматический словарь древнечешского с учетом звуковых изменений (таких, как палатализации), чередования основ и изменения типов склонения. Преимущество словарных таггеров проявляется в разборе слов с чередованиями на стыке основы и окончания, а также слов нерегулярных словоизменительных типов. Однако если такой анализатор встречает слово, отсутствующее в словаре, качество разбора падает, даже если система снабжена модулем порождения гипотез для незнакомых слов[9].

Рынок статистических разметчиков стал активно развиваться с появлением больших корпусов, морфологически размеченных вручную. Так, хорошее качество разметки для современного русского языка показывают, например, системы Mocky и ODA, обученные на 6-миллионном корпусе НКРЯ со снятой омонимией[10]. Примером удачного применения статистического таггера можно считать морфологическую разметку корпуса средневекового португальского языка[11]. К текстам, предварительно снабженным частеречной разметкой, применялся статистический анализатор, рассчитанный на современный португальский язык и натренированный на небольшом объеме средневековой португальской лексики. Используя статистику реализации окончаний в базе n-грамм, т. е. цепочек из двух, трех и более идущих подряд словоформ, статистические таггеры, как правило, показывают очень высокое качество частеречной аннотации, в том числе и для редких слов. Вместе с тем качество определения лемм непредсказуемо, в частности для словоформ с чередованиями могут порождаться леммы, отсутствующие в языке в принципе.

Помимо словарного и статистического подходов существует еще один способ разметить исторические тексты — копирование разметки параллельно выравненного перевода исторического текста на современный язык. Подобная технология была применена для разметки Библии, написанной на среднеанглийском языке[12], а также для снятия омонимии в тексте Повести Временных Лет,

[9] См.: *Ляшевская О., Астафьева И., Бонч-Осмоловская А., Гарейшина А., Гришина Ю., Дьячков В., Ионов М., Королева А., Кудринский М., Литягина А., Лучина Е., Сидорова Е., Толдова С., Савчук С., Коваль С.* Оценка методов автоматического анализа текста: морфологические парсеры русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (2010). Вып. 9 (16). М.: РГГУ, 2010. С. 318–326.

[10] См.: *Sharoff S., Nivre J.* The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (2011). Вып. 10 (17). М.: РГГУ, 2011.

[11] См.: *Rocio V., Alves M. A., Lopes J. G., Xavier M. F., Vicente G.* 1999. Automated creation of a partially syntactically annotated corpus of Medieval Portuguese using contemporary Portuguese resources. Proceedings of the ATALA workshop on Treebanks. Paris. В настоящее время разрабатывается и таггер для древнерусского и среднерусского языка, см. *Berdičevskis A., Eckhoff H. M., Gavrilova T.* The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» 2016. Вып. 15 (22) (forthc); о первых результатах его работы см. в разделе 8.

[12] См.: *Moon T., Baldridge J.* Part-of-speech tagging for middle English through alignment and projection of parallel diachronic texts. Proceedings of the 2007 Joint Conference on Empirical Me-

написанной на древнерусском языке и предварительно проанализированной с помощью статистического таггера[13]. Данный метод анализа текстов позволяет добиться высоких результатов, если разница между древним и современным состоянием языка не очень велика, а выравнивание произведено качественно. Тем не менее данная технология может быть применена только в отношении текстов, переведенных на современный язык, что делает невозможным ее применение для основной массы старорусских текстов НКРЯ.

В данной работе в качестве основы для словарного анализатора мы использовали Юни-таггер Т. А. Архангельского[14]. Эта программа предназначена для разметки разных и разноструктурных языков и требует на вход два файла: словарь обрабатываемого языка и список парадигм всех изменяемых частей речи этого языка, оба в специальном формате. Словарь должен включать в себя лемму слова, его главную и все косвенные основы, а также ссылку на парадигму того словоизменительного типа, к которому оно принадлежит. Факультативно может быть добавлен перевод. В парадигмах должны быть представлены изменяющиеся части всех словоизменительных типов данного языка.

Парадигма	N3t
Примеры	жен-а
ед. им.	жен-а
ед. вин.	жен-у
ед. род.	жен-ы
...	...

Поскольку в цифровом виде ни одного древнерусского словаря, в котором были бы указаны словоизменительные типы лексем, не существует, за основу был взят электронный словарь А. Е. Полякова, ранее разработанный для разметки церковнославянских текстов НКРЯ и ориентированный, соответственно, на словоизменительные парадигмы церковнославянского языка[15].

Под грамматической парадигмой подразумевается список словоизменительных морфем (преимущественно флексий), соотнесенных с грамматическими формами, которые они маркируют.

Например, элемент парадигмы для словоизменительного типа «существительные *а*-склонения с твердым согласным на конце основы без чередований» (ср. *жена*) будет записан следующим образом:

---

thods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, June 28–30, 2007. С. 390–399.

[13] См.: *Meyer R.* New wine in old wineskins? Tagging Old Russian via annotation projection from modern translations. Russian linguistics, 35 (2), 2011. С. 267–281.

[14] См.: *Архангельский Т. А.* Принципы построения морфологического парсера для разноструктурных языков. Дисс... канд. филол. наук. М.: МГУ, 2012.

[15] Электронный грамматический словарь размещен в открытом доступе на сайте http://feb-web.ru/febupd/slavonic/dicgram. См.: *Поляков А. Е., Савчук С. О., Сичинава Д. В.* Грамматический словарь для автоматического анализа текстов XVIII–XIX веков: первые результаты // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 12 (19). М.: Изд-во РГГУ, 2013. С. 633–654.

paradigm: N3t
flex: <0>.ы
gramm: sg,gen

Здесь N3t — условное обозначение словоизменительного типа, в поле flex указано окончание *ы*, присоединяемое к главной, здесь единственной основе (<0>), а в поле gramm — грамматические характеристики формы (единственное число, родительный падеж: *жен-ы*).

Каждой лексеме в словаре приписано название парадигмы, содержащей все возможные для этого слова флексии. Для того чтобы на таком материале построить анализатор, способный отделять не только церковнославянские, но и древнерусские флексии, необходимо было составить древнерусские парадигмы, аналогичные парадигмам А. Е. Полякова. Это значит, что для правильного анализа слова *рабъ*, соотнесенного в словаре со словоизменительным типом N1t, должна существовать не только парадигма N1t для церковнославянского, но и аналогичная парадигма, содержащая древнерусские окончания. Мы вынуждены сделать сильное допущение, что деление лексики на словоизменительные типы в древнерусском языке в достаточной мере соотносится с делением лексики в церковнославянском. Иными словами, если для лексемы X с церковнославянской парадигмой P была создана древнерусская парадигма P1, то предполагается, что и лексема Y с той же церковнославянской парадигмой P в древнерусском языке будет изменяться по парадигме P1. Это предположение относительно условно, но спровоцировано отсутствием цифрового древнерусского словаря, в котором были бы размечены словоизменительные типы.

Так как многие флексии древнерусского языка по форме совпадают с соответствующими флексиями церковнославянского, мы объединили вместе соответствующие друг другу парадигмы церковнославянского и древнерусского языков. Это было сделано, чтобы упростить разметку и не создавать слишком много вариантов анализа одного и того же слова.

Далее мы опишем основные этапы расширения и адаптации грамматического словаря к разметке данных среднерусского корпуса.

4. Пополнение грамматического словаря

4.1. Автоматическое порождение косвенных основ из грамматического словаря церковнославянского языка

Словарные статьи словаря церковнославянского языка не содержат косвенных основ, поскольку их исчисление «вшито» в программу морфологического анализа А. Е. Полякова. Однако принцип разметчика Юни-парсер другой: и в парадигмах, и в словаре анализатора должны быть указаны косвенные основы. Следовательно, необходимо было составить правила, порождающие косвенные основы для каждого словоизменительного типа, в котором они встречаются. Ниже приведен фрагмент представления парадигмы N1k* (тип *свиток*), где <0>,<1> — ссылки на тип основы:

paradigm: N1k*

flex: <1>.ома
gramm: du,ins
flex: <0>.ъ//<0>.
gramm: sg,nom
flex: <1>.у
gramm: du,loc

Здесь указано, что именная парадигма N1, подтип k* (основа, заканчивающаяся на *-к* с беглым гласным перед ним), имеет в твор. падеже двойств. числа окончание *-ома*, которое присоединяется к основе <1>; в форме им. падежа ед. числа к основе <0> присоединяется *-ъ* или «ничего» (нулевое окончание); в форме местн. падежа двойств. числа к основе <1> присоединяется окончание *-у*.

Набор правил специфичен для парадигмы, но их можно обобщить в следующий список:

А) вставка или удаление гласного в конце основы;

Б) мена гласного в конце основы;

В) вставка или удаление согласного в конце основы;

Г) мена согласного или кластера.

А. Вставка или удаление гласного. Это правило характерно скорее для именных парадигм, ср. *оселъ* (им. ед.) — *осла* (род. ед.). Тем не менее оно присутствует и в некоторых глагольных парадигмах, например в парадигме V14t* (тип *чисти — чьту*). В зависимости от конкретного слова может вставляться / выпадать либо *о*, либо *е* (в парадигме V11e (тип *краснѣти*) — ѣ). В некоторых случаях, когда определить вставной гласный автоматически по форме слова невозможно (ср. *бревно / бревенъ* и *окно / оконъ*, в обоих случаях основа им. падежа заканчивается на кластер твердых согласных), правило было устроено так, чтобы порождать два варианта одной основы. К правилам этого типа можно также отнести правила, порождающие косвенные основы для глаголов парадигмы V15ol: *кла- / кол-* (*клати — колю*) или парадигмы V15or: *бра- / бер-* (*брати — беру*).

Б. Мена гласного. В некоторых глагольных парадигмах основа с гласным меняется на основу с согласным в той же позиции: *мя- / мн-* (*мяти*, V15n). Существуют также случаи замены гласного гласным: *пѣ- / пой-* (*пѣти*, V15e).

В. Вставка или удаление согласного. Наиболее частым добавляемым согласным является *л*, ср. парадигмы V21p (*ѡбъяви́ти*), V12p (*колеба́ти*), V13p+V22p (*спа́ти*). Кроме того, в косвенной основе могут появляться и другие согласные, например *б, п* (*погребсти́*, V14p, *почерпсти́*, V14p — инновационные формы вместо *погрети*, *почрѣти*), *т* (*помести́*, V14t), *д* (*вести́*, V14d), *в* (*да́ти*, V15v).

Г. Мена согласного или консонантного кластера. Двумя наиболее частыми случаями в этом типе чередований являются изменения заднеязычных согласных перед окончаниями, начинающимися на передний гласный, характерные как для именных, так и для глагольных основ, а также замены зубных согласных на шипящие (т. е. проявления 1-й и 2-й палатализации), ср.:

к-(ц)-ч: отрок-ъ (им. ед.) — *отроц-ѣ* (местн. ед.) — *отроч-е/ъ* (зват. ед.)
г-(з)-ж: враг-ъ (им. ед.) — *враз-ѣ* (местн. ед.) — *враж-е/ъ* (зват. ед.)
*х-(с)-ш: дух-ъ (*им. ед.) — *дус-ѣ* (местн. ед.) — *душ-е/ъ* (зват. ед.)

Парадигмы V21t (*роди́ти*), V22t (*ненави́дѣти*), V12t (*страда́ти*) и V12t+V22t (*восхотѣ́ти*) характеризуются чередованиями *-д- / -жд- (родити — рожду*), *-з- / -ж- (возити — вожу*), *-с- / -ш-* (*висѣти — вишу*), *-т- / -щ- (восхотети — восхощу*), *-ст- / -щ-* (*растити — ращу*). Существуют и другие, менее частотные, чередования, такие как чередование *-ст- / -т-* в парадигме V14st (*изѡбрѣсти́*).

В некоторых основах происходит варьирование кластера, как правило многобуквенного суффикса или суффиксов. Такова парадигма V12ov (*требовати*), где суффикс *-ов- / -ев-* исчезает в большинстве других форм.

Для части глаголов косвенные основы были добавлены в словарь вручную, поскольку они не порождались по правилам, общим для всех глаголов их словоизменительного типа. По этой же причине все глаголы парадигмы Viti (глагол *идти* и глаголы, образованные от него при помощи приставок) были также записаны в словарь вручную (ср. *ид-/и-/ш-/ше-/шед- с* ***в*****н***ид-/***в*****н***и-/***в*****о***ш-/***в*****о***ше-/***в*****(о)***шед-).

4.2. Добавление новых типов склонения, вызванных историческими переходами

Склонения раннедревнерусского периода претерпели существенные изменения, со временем перейдя в три современных склонения. Так, 1-е современное склонение (*жена*) было образовано из старого женского *а*-склонения, 2-е современное (*двор*) — из *о-*, *а-*, *и-*, *-i* склонений мужского рода и т. д. Наш анализатор ориентирован на разбор текстов большого временного промежутка, предположительно включающего в себя и древнее и современное состояние системы склонений, поэтому словам, которые диахронически изменили свое склонение, были приписаны сразу два типа — новый и старый. Например, слова из парадигмы N6t (*рабъ*) получили ссылки на словоизменительные типы N1t, так как все имена мужского рода *о*-склонения (им. мн. *раб-и*) перешли в а-склонение (им. мн. *раб-ы*). При составлении парадигм были учтены также такие важные диахронические изменения, как, например, отмена палатализации в первом склонении (*враз-и / враг-и*) или унификация окончаний *-ам(ъ), -ами, -ах(ъ)* для дательного, творительного и местного падежей множественного числа всех склонений.

4.3. Предсказание парадигмы и порождение основ для глаголов с неполной грамматической аннотацией в церковнославянском словаре

Для глаголов, обладающих неполной грамматической аннотацией в церковнославянском словаре, парадигма порождалась автоматически исходя из внешнего вида леммы. Были применены следующие правила:

- Для глаголов, оканчивающихся на *-ити* или *-ѣти,* были добавлены возможные парадигмы V21n, V21a, V21s, V21p, V21t, V22n, V22p, V22t, V22s, V22a (ср. *творити, строити, рѣшити, любити, родити, звенѣти, терпѣти, видѣти, слышати, стояти*).
- Для глаголов, оканчивающихся на *-ати* или *-ѣти,* были добавлены возможные парадигмы V11a, V11e, V12ov, V12n, V12p, V12t, V12k, V12a, V12x, V12x* (ср. *делати, краснѣти, требовати, глаголати, сыпати, страдати, алкати, сѣяти, рвати, звати*).
- Для глаголов, оканчивающихся на *-нути,* были добавлены возможные парадигмы V13a, V13t, V13k (ср. *минути, гибнути, двигнути*).

- Для глаголов, оканчивающихся на *-щи,* были добавлены возможные парадигмы V14k, V14g, V14g*, V14eg (ср. *пещи, мощи, жещи, лещи*).
- Для глаголов, оканчивающихся на *-сти,* были добавлены возможные парадигмы V14p, V14z, V14t, V14d, V14st, V14t*, V14ed (ср. *грести, нести, месть, красти, расти, чисти, състи*).
- Для глаголов, оканчивающихся на *-ити,* была добавлена парадигма V15i (*бити*).
- Для глаголов, оканчивающихся на *-ыти,* была добавлена парадигма V15y (*мыти*).
- Для глаголов, не оканчивающихся ни на одно из вышеперечисленных буквосочетаний, были добавлены возможные парадигмы V15e (*пѣти*), V15n (*мяти*), V15a (*стати*), V15v (*жити*).
- Глаголам, четвертой буквой от конца в которых была *л,* были добавлены парадигмы V15ol (*клати — колю*), V15el (*млети — мелю*), а глаголам, четвертой буквой от конца в которых была *р,* — парадигмы V15or (*брати — борю*) и V15er (*умрети — умру*).

Затем для каждой из предложенных парадигм были автоматически порождены основы таким образом, что лексема с каждой парадигмой и набором основ представляла отдельный словарный вход. Если для автоматически предложенной парадигмы не удавалось породить косвенные основы (то есть правила их порождения были неприменимы к данному глаголу), лексема с такой парадигмой удалялась.

4.4. Обработка нерегулярных глаголов и глаголов, обладающих особым типом спряжения

Косвенные основы некоторых глаголов было невозможно породить автоматически, используя правила, описанные в разделе 4.3. Для таких глаголов был вручную прописан образец изменения основы, а затем для них и всех однокоренных им глаголов основы порождались автоматически согласно заданному образцу вне зависимости от парадигмы, к которой они принадлежат. Список таких «нерегулярных» глаголов и их основ (в нормализованной орфографии) приведен ниже:

имати / им / емл; *гнати / гн / жен*; *слати / сл / шл*; *звати / зв / зов*; *ткати / тк / ток/тек*; *ржати / рж / рож/реж*; *ссати / сс / сос*; *рещи / ре / рек/рк / реч/рч / рец/рц; блещи / бле/бол / блек/болк / блеч/болч / блец/блоц.*

Отдельного внимания заслуживает глагол *яти*. Для этого глагола без приставки существует всего две косвенные основы — *им* и *ем*. Однако с прибавлением приставки количество косвенных основ увеличивается и может достигать семи. Существуют следующие основные типы изменения основ глаголов с корнем *я* в зависимости от приставки:

• *приять*	*прия/приня*	*приим/прим/прием/прийм/приним/принем*
• *отъять*	*отъя/отня*	*отъим/отъем/отым/отьим/отним/отнем/отьним*
• *сжати*	*сжа*	*сжим/сожм/сжем*

Для глаголов с корнем *я* все варианты основ для каждой приставки были прописаны вручную.

Наконец, для глаголов уникальных типов спряжения — *идти, ехати, ведати, быти* и однокоренных — также был указан образец изменения основы, по которому порождались основы для них и всех однокоренных глаголов.

Так как глагол *быти* является супплетивным, для него была разработана особая стратегия разметки. Все формы глагола *быти* были записаны в словарь флексий, а возможные приставки — как основы в грамматический словарь. Например, слово *отбывахом* было бы разобрано как *от-бывахом*.

Отдельную сложность представляет глагол *идти*. Его корень, как и корень глагола *яти*, изменяется по-разному в зависимости от типа приставки, к тому же возможны такие варианты основы, как *ийти*, *ити* и *итти*. Для глаголов, однокоренных глаголу *идти,* косвенные основы порождались согласно следующим правилам:

- если приставка заканчивается на гласный или приставки нет, то по образцу *ид-и-ш-ше-шед*;
- если приставка в инфинитиве заканчивается на *н*, то по образцу *внид-вни-вош-воше-в(о)шед*;
- если приставка в инфинитиве заканчивается на другой согласный (не *н*), то по образцу *изъид/изыд-изъи/изы-изше-изшед*.

5. Альтернативный подход, связанный с использованием существующих ресурсов

В предыдущих разделах были описаны этапы создания системы лексико-грамматической аннотации старорусских текстов, основанной на словарно-правиловом подходе. В этом разделе мы рассмотрим второй, упрощенный подход, связанный с привлечением существующих ресурсов. Его идея заключается в разметке текстов корпуса таггерами Mystem и TreeTagger[16], предназначенными для современных русских текстов, и использовании множества разборов древнерусского корпуса НКРЯ. В этом случае также не ставилась задача разрешения неоднозначности разборов (дизамбигуации) или минимизации их количества, поэтому критерий «широкого охвата» предусматривал, что порождаемое множество разборов считается правильным, если правилен хотя бы один из предложенных разборов. Наш подход строится на простом допущении, что современные анализаторы справятся с разбором словоформ, соответствующих современным, а древнерусские разборы покроют достаточно большую часть «старого пласта» словоформ, в том числе наиболее частотные формы, такие как формы глагола

[16] Использовалась программа Mystem, адаптированная для НКРЯ. О принципах работы Mystem см.: *Segalovich I.* 2003. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. Proceedings of MLMTA, Las Vegas, Nevada. P. 273–280. Версия TreeTagger для современного русского языка была обучена на корпусе со снятой лексико-грамматической омонимией НКРЯ (6,5 млн. словоупотреблений) А. Феногеновой, О. Дерезой и Д. Каютенко. URL: http://web-corpora.net/wsgi/mystemplus.wsgi/mystemplus/mystem/.

быть (*бе, бѣаху, бѣаше, бѣше, быста, быхъ* и т. д.) или местоимения *иже* (*егоже, еже, немже* и т.д.).

Так же, как и словарный подход, обсуждаемый в этом разделе гибридный разметчик включает модуль нормализации орфографии текста, а именно приведение ее к позднерусскому орфографическому узусу[17]. Кроме того, был реализован экспериментальный модуль преобразования графических окончаний, который в разных вариантах должен был приводить к современной норме отдельные грамматические формы, такие как инфинитив глагола (*бояти*), 2 л. ед. числа наст.-буд. времени (*будеши*), мн. числа пассивных причастий (*украшени*), ж. род ед. числа имен на *-ий*, *-ие* (житиа), приписывать клитике *ся* (которая в старорусском корпусе считается отдельным токеном) помету «частица», а также склеивать лемму глагола и следующую за ней клитику *ся* в одну лемму. Применение правил изменения окончаний в этих случаях может приносить как положительный эффект (увеличение зоны покрытия словника разборами), так и отрицательный (ошибочные написания вида *уношь* вместо *уноши*, *выть* вместо (*с*) *выти* ‘с участка земли’ и т. п.), поэтому при измерении оценки качества работы разметчика (см. следующий раздел) мы запускали этот модуль с разными настройками, чтобы оценить его эффективность.

6. Сравнение результатов работы автоматических разметчиков

Для оценки качества распознавания форм словоизменения мы использовали два текста конца XVI — середины XVII в.: «Житие Сергия Радонежского», воспроизводящее более ранний список, составленный в 1417–1418 гг., и «Наказ Афанасию Филипповичу Пашкову на воеводство в Даурской земле». Первый документ размечен вручную разработчиками корпуса TOROT[18]; второй документ был размечен по тегам части речи и леммы одним из авторов статьи. Объем текстов составляет 19920 и 12769 словоупотреблений соответственно.

Оценивалось качество распознавания части речи и леммы. Оценка качества определения грамматических признаков пока не проводилась в связи с ее трудоемкостью. Как уже было сказано, разметка словоформы признавалась удовлетворительной, если хотя бы один разбор совпадал с разбором в ручной разметке («мягкий» принцип). При этом частеречные теги стандарта TOROT были переведены в теги древнерусского корпуса НКРЯ.

В таблицах 1 и 2 приведены результаты оценки качества распознавания части речи. Полнота оценивалась как количество выданных системой разборов, поделенное на общее число слов в тесте (R). Аккуратность оценивалась как количество удовлетворительных разборов, поделенное на количество всех словоформ в тестовом корпусе (A_{soft}[19]). Точность оценивалась как количество удовлет-

[17] См. предисловие к Словарю русского языка XI–XVII вв. Вып. 30. М.: Наука; Азбуковник, 2015.

[18] Текст с разметкой выложен на сайте https://nestor.uit.no/sources/215. Мы благодарим разработчиков корпуса Х. М. Экхоф и А. Бердичевского за предоставленные материалы по оценке качества морфологической разметки и правила конверсии частеречных тегов.

[19] Здесь «soft» означает «мягкий» принцип оценки, см. выше. Аккуратность и точность системы, выдающей 100 % ответов, по определению равны.

ворительных разборов, поделенное на общее количество выданных системой разборов (P_{soft})[20].

Таблица 1

Качество распознавания части речи в «Житии»

Тип таггера	Полнота (R)	Аккуратность (A_{soft})	Точность (P_{soft})
Словарный	80,6 %	76,3 %	94,7 %
Гибридный	100 %	93,8 %	93,8 %

Таблица 2

Качество распознавания части речи в «Наказе»

Тип таггера	Полнота (R)	Аккуратность (A_{soft})	Точность (P_{soft})
Словарный	70,8 %	68,7 %	97,1 %
Гибридный	100 %	95,7 %	95,7 %

Вполне ожидаемо, что словарный таггер показывает меньшую полноту, чем гибридный, который благодаря системе порождения гипотез для несловарных слов покрывает тестовый корпус на 100 %. Полнота словарного разметчика на тексте «Наказа» составляет всего 70,8 %, что значительно ниже, чем тот же показатель на тексте «Жития» (80,6 %). Это связано с повышением доли редких слов и имен собственных в первом из упомянутых тестовых фрагментов. Точность разметки гибридного таггера колеблется в районе 94 %, однако если бы этот таггер не учитывал разборы древнерусского корпуса, его точность была бы на 12–13 % ниже. Качество работы словарного разметчика в тех случаях, где он выдает ответ, ожидаемо выше, чем качество гибридного (см. показатели точности Psoft), однако в целом гибридный разметчик выдает больше правильных разборов (см. показатели аккуратности Asoft).

Таблицы 3 и 4 показывают результаты сопоставления качества лемматизации, сделанной словарным таггером и гибридным анализатором. Оценка лемматизации проводилась независимо от того, правильно разобрана часть речи или нет. Впрочем, случаи, когда лемма распознана правильно, а часть речи — нет, немногочисленны (1–3 % всех словоформ тестового корпуса).

[20] Токены пунктуации не учитывались в подсчетах. Обратим внимание, что учет токенов пунктуации (как это принято в методиках оценки некоторых таггеров, см., например: *Sharoff S., Nivre J.* The proper place of men and machines in language technology processing Russian without any linguistic knowledge // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (2011). Вып. 10 (17). М.: РГГУ, 2011) может увеличивать показатели на 5–10%.

Таблица 3

Качество определения леммы в «Житии»

Тип таггера	Полнота (R)	Аккуратность (A_{soft})	Точность (P_{soft})
Словарный	80,6 %	76,1 %	94,5 %
Гибридный	100 %	89,5 %	89,5 %

Таблица 4

Качество определения леммы в «Наказе»

Тип таггера	Полнота (R)	Аккуратность (A_{soft})	Точность (P_{soft})
Словарный	70,8 %	68,8 %	97,2 %
Гибридный	100 %	92,1 %	92,1 %

Как видно, точность словарного таггера в части определения леммы составляет 94,5–97,2 %. Точность гибридного таггера на тексте «Наказа» также выше 90 %, однако на данных «Жития» она падает до 89,5 %. В целом можно сделать вывод, что более «современный» текст «Наказа», с меньшим количеством архаических форм, разбирается лучше гибридным разметчиком, тогда как текст «Жития» относительно точнее разбирает словарный разметчик.

Таким образом, по нашим оценкам, качество определения части речи и леммы обоих лексико-грамматических анализаторов составляет 89,5% и выше, что можно признать удовлетворительным результатом для организации так называемого неточного поиска по корпусу, а также для передачи разметки для последующего ручного постредактирования. Конечно, при этом требуется помнить, что в тестовый корпус вошли только два текста и на других текстах, с большим количеством архаичных и нерегулярных форм, оценки могут быть менее оптимистичными.

Во время подготовки статьи к публикации появились также первые итоги тестирования статистического таггера[21], натренированного на разметке корпуса TOROT: при несколько иной методике оценки и на несколько ином объеме тестовых данных он дает 89,5 % аккуратности распознавания части речи и 69,8 % аккуратности распознавания комбинации части речи и леммы. В комбинации со словарным таггером его аккуратность возрастает до 91,4 % и 78,5 % соответственно.

Дальнейшие перспективы развития лексико-грамматических анализаторов для старорусских текстов видятся в комбинации всех трех подходов. Так, статистический таггер наиболее эффективен для определения частеречных тегов, в то время как расширение грамматического словаря может помочь в уточнении

[21] См.: *Berdičevskis A., Eckhoff H. M., Gavrilova T.* The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» . Вып. 15 (22) (forthc.)

леммы и части речи более редких случаев и нестандартных форм. Определенный прирост может дать автоматизация построения словаря с использованием методик объединения словоформ в кластеры, относящиеся к одной парадигме[22]. Вместе с тем требуется большая работа по установлению соответствий между разборами древнерусского, церковнославянского корпуса НКРЯ, корпуса берестяных грамот и корпуса TOROT и других исторических корпусов русского языка, а также леммами исторических словарей, если мы хотим использовать все доступные источники для улучшения качества анализа.

Ключевые слова: старорусский язык, корпус, НКРЯ, лексико-грамматическая разметка, морфологический таггер, грамматический словарь, именное словоизменение, глагольное словоизменение.

Lexico-Grammatical Annotation of the Middle Russian Corpus 1400–1700: A Computational Approach

T. Gavrilova, T. Shalganova, O. Liashevskaia

The paper discusses two approaches to the automatic lexico-grammatical tagging of the Middle Russian texts (1400–1700), included in the Russian National Corpus (RNC). The task is to assign each token a part of speech label, a tuple of grammatical features, and a lemma (without disambiguation). Middle Russian combines, on the one hand, features of the earlier state of the grammatical system, including aorist and imperfect verb forms, the dual number, a number of archaic inflectional paradigms, and, on the other hand, features of modern Russian inflectional morphology. In lexicon, we can see the same mix of Old Russian and Modern Russian lemmas. Moreover, the texts can contain Church Slavonic and dialectal forms. Absence of a standardised orthography and absence of a standard variant pose even more challenges to processing Middle Russian texts.

The first approach is based on writing an electronic dictionary of Old Russian and building a module to handle spelling inconsistency. In the absence of open electronic resources for Middle Russian morphology, an electronic dictionary of Church Slavonic

[22] См.: *Ляшевская О. Н., Сичинава Д. В., Кобрицов Б. П.* Автоматизация построения словаря на материале массива несловарных словоформ // Интернет-математика — 2007 / П. И. Браславский, отв. ред. Екатеринбург: Изд-во Урал. ун-та, 2007. С. 118–125; *Клышинский Э. С.* Некоторые сложности автоматизированной лемматизации несловарных словоформ // Материалы международной конференции «Диалог 2009». М.: РГГУ, 2009. С. 165–169; *Сокирко А. В.* Быстрословарь: предсказание морфологии русских слов с использованием больших лингвистических ресурсов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 9 (16). М.: Изд-во РГГУ, 2010. С. 450–456.

was expanded and adapted to Middle Russian. The paper describes the steps required to change nominal and verbal entries in this dictionary. We follow the principle of «a wider expansion» which presupposes that the analyser is allowed to generate as many annotations as possible so that at least one annotation would be correct.

The second approach uses, firstly, an existing Modern Russian tagger supplemented by the module reducing spelling variation, and secondly, a database of lexico-grammatical annotations retrieved from the Diachronic corpus of the RNC.

We evaluate the output of both analysers against a manually annotated test data. We also discuss the benchmark scores and outline future prospects for the development of the Middle Russian taggers.

Keywords: Middle Russian, Russian National Corpus, lexico-grammatical tagging, morphological analysis, grammatical dictionary, spelling variation, nominal inflection, verb inflection.

Список литературы

1. Историческая грамматика русского языка: морфология; глагол /[Р. И. Аванесов, В. В. Иванов, В. Б. Силина и др.] *Р. И Аванесов, В. В. Иванов*, ред. М.: Наука. 1982.
2. *Архангельский Т. А.* Принципы построения морфологического парсера для разноструктурных языков. Дисс... канд филол. наук. М.: МГУ, 2012.
3. *Демьянов В. Г.* Вести-Куранты: 1. Издание для исследования. 2. Исследование для издания // Лингвистическое источниковедение и история русского языка. М., 2000. С. 213–232.
4. *Добрушина Е. Р., Кравецкий А. Г., Поляков А. Е.* Корпус и частотный грамматический корпусный словарь церковнославянского языка в составе НКРЯ // Труды Института русского языка им. В. В. Виноградова. Вып. 6. 2015. С. 116–141.
5. *Добрушина Е.Р., Поляков А. Е.* Корпус церковнославянского языка: возможности, методы создания, перспективы // Вестник ПСТГУ. Серия III: Филология. 2013. Вып. 1 (31). С. 32–44.
6. *Живов В. М.* Очерки исторической морфологии русского языка XVII–XVIII веков. М., 2004.
7. *Зализняк А. А.* Грамматический словарь русского языка: Словоизменение. М.: Русский язык, 1977. 4-е изд., испр. и доп. — М.: Русские словари, 2003.
8. *Зобнин А. И., Пичхадзе А. А.* Корпус древнерусских переводов XI–XII вв.: результаты и перспективы // Научно-техническая информация. Серия 2: Информационные процессы и системы. № 3. 2005. С. 44–47.
9. *Клышинский Э. С.* Некоторые сложности автоматизированной лемматизации несловарных словоформ // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 8 (15). М.: РГГУ. 2009. С. 165–169.
10. *Кривко Р. Н.* (гл. ред.). Словарь русского языка XI–XVII вв. Вып. 30 (Томъ — Уберечися). М.: Наука — Азбуковник, 2015.
11. Историческая грамматика древнерусского языка */В. Б Крысько,* ред. Т. 1–4. М.: Азбуковник, 2000–2006.
12. *Ляшевская О., Астафьева И., Бонч-Осмоловская А., Гарейшина А., Гришина Ю., Дьячков В., Ионов М., Королева А., Кудринский М., Литягина А., Лучина Е., Сидорова Е., Толдова С., Савчук С., Коваль С.* Оценка методов автоматического анализа текста: морфологические парсеры русского языка // Компьютерная лингвистика и интеллектуаль-

ные технологии: По материалам ежегодной Международной конференции «Диалог» (2010). Вып. 9 (16). 2010. М.: РГГУ. С. 318–326.
13. *Ляшевская О. Н., Плунгян В. А., Сичинава Д. В.* О морфологическом стандарте Корпуса современного русского языка // Национальный корпус русского языка: 2003–2005. М.: Индри, 2005. С. 111–135.
14. *Ляшевская О. Н., Сичинава Д. В., Кобрицов Б. П.* Автоматизация построения словаря на материале массива несловарных словоформ // Интернет-математика — 2007: Сборник работ участников конкурса научных проектов по информационному поиску / П. И Браславский., отв. ред. Екатеринбург, 2007. С. 118–125.
15. *Мишина Е. И., Пичхадзе А. А.* Древнерусский подкорпус Национального корпуса русского языка // Труды Института русского языка им. В. В. Виноградова РАН. Вып. 6. 2015. С. 99–115.
16. *Молдован А. М.* Памятники древнерусской письменности в Национальном корпусе русского языка // Труды Института русского языка им. В. В. Виноградова РАН. Вып. 6. 2015. С. 88–98.
17. *Пичхадзе А. А.* Корпус древнерусских переводов XI–XII вв. и изучение переводной книжности Древней Руси // Национальный корпус русского языка: 2003–2005. М., 2005. С. 251–262.
18. *Поляков А. Е.* Грамматический словарь церковнославянского языка (по материалам корпуса). URL: http://feb-web.ru/febupd/slavonic/dicgram.
19. *Поляков А. Е.* Проблемы и методы анализа русских текстов в дореформенной орфографии // Компьютерная лингвистика и интеллектуальные технологии: По материалам Международной конференции «Диалог 2012». Вып. 11 (18). М.: Изд-во РГГУ, 2012. С. 536–547.
20. *Поляков А. Е.* Корпус церковнославянских текстов в составе Национального корпуса русского языка, первая версия: проблемы и решения // Доклад на международной научной конференции «Информационные технологии и письменное наследие (El'Manuscript-12)». Петрозаводск, 2012.
21. *Поляков А. Е.* Корпус церковнославянских текстов: проблемы орфографии и графики // Przegląd wschodnioeuropejski V/1. 2014. С. 245–254.
22. *Поляков А. Е., Савчук С. О., Сичинава Д. В.* Грамматический словарь для автоматического анализа текстов XVIII–XIX веков: первые результаты // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 12 (19). М.: Изд-во РГГУ, 2013. С. 633–654.
23. *Сичинава Д. В.* Исторические корпуса Национального корпуса русского языка как инструмент диахронических исследований грамматики // Писменото наследство и информационните технологии: Материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / В. А. Баранов, В. Желязкова, А. М. Лаврентьев, отв. ред. София; Ижевск, 2014.
24. *Сокирко А. В.* Быстрословарь: предсказание морфологии русских слов с использованием больших лингвистических ресурсов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 9 (16). М.: Изд-во РГГУ, 2010. С. 450–456.
25. *Успенский Б. А.* История русского литературного языка (IX–XVII вв.). М., 2002. *Berdichevskis A., Eckhoff H. M., Gavrilova T.* Forthcoming. The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» 2016. Вып. 15 (22) (forthc.)
26. *Jínová P., Lehečka B., Oliva K.* Describing Old Czech Declension Patterns for Automatic Text Analysis // Mundo Eslavo. № 13. 2014. P. 7–17.

27. *Meyer R.* Semi-automatic morphosyntactic tagging of a diachronic corpus of Russian // Proceedings of the Corpus Linguistics Conference, CL2009 / Mahlberg M., González-Díaz V., Smith C., eds. Liverpool, 2009. P. 20–23.
28. *Meyer R.* New wine in old wineskins? Tagging Old Russian via annotation projection from modern translations // Russian linguistics. № 35 (2). 2011. P. 267–281.
29. *Moon T., Baldridge J.* Part-of-speech tagging for middle English through alignment and projection of parallel diachronic texts // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Respublic, June 28–30. 2007. P. 390–399.
30. *Rocio V., Alves M. A., Lopes J. G., Xavier M. F., Vicente G.* Automated creation of a partially syntactically annotated corpus of Medieval Portuguese using contemporary Portuguese resources // Proceedings of the ATALA workshop on Treebanks. Paris, 1999.
31. *Segalovich I.* A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engin //Proceedings of MLMTA, Las Vegas, Nevada, 2003. P. 273–280.
32. *Sharoff S., Nivre J.* The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (2011). Вып. 10 (17). М.: РГГУ, 2011.
33. *Sporleder C.* Natural language processing for cultural heritage domains // Language and Linguistics Compass. 4. 9. 2009. P. 750–768.