Вестник ПСТГУ.
Серия III: Филология.
2017. Вып. 51. С. 11–20

Гаврилова Татьяна Сергеевна,
Национальный исследовательский
университет «Высшая школа экономики»
101000, Россия, Москва, Мясницкая ул., д. 20
tanya96gavrilova@yandex.ru

Шалганова Татьяна Александровна,
Национальный исследовательский
университет «Высшая школа экономики»
101000, Россия, Москва, Мясницкая ул., д. 20
insana7@mail.ru

Ляшевская Ольга Николаевна,
канд. филол. наук, Национальный исследовательский
университет «Высшая школа экономики»,
Институт русского языка им. В. В. Виноградова РАН
101000, Россия, Москва, Мясницкая ул., д. 20
olesar@yandex.ru

Взіалъ, възялъ, вьзял: обработка орфографической вариативности при лексико-грамматической аннотации Старорусского корпуса XV–XVII вв.*

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

Рассматривается проблематика нестабильной орфографии корпуса текстов позднедревнерусского периода в свете их автоматической обработки. Тексты Старорусского корпуса Национального корпуса русского языка (НКРЯ) включают памятники, написанные преимущественно в XV–XVII вв., т. е. в тот период, когда вариативность написания слов все еще была нормой. Задача лексико-грамматической разметки словоформ в корпусе заключается в определении начальной формы (словарной формы, леммы), части речи и грамматических характеристик. Традиционные методы автоматического определения лексико-грамматических характеристик базируются на презумпции идентичного вида основы и окончания слова в каждой из грамматических форм. Поэтому нестабильная орфография памятников становится причиной неэффективной работы автоматических морфологических анализаторов (таггеров) — в том случае, если они не оснащены модулем поддержки орфографической вариативности. В работе применяется относительная и абсолютная нормализации орфографии. Относительная нормализация предполагает размножение орфографических представлений основ и окончаний в грамматическом словаре по регулярным правилам, обрабатывающим: а) флексии; б) именные основы с регулярной вариативностью *-ск*(*ии*) / *ст*(*ии*), *-и*(*я*) / *-ь*(*я*); в) основы имен церковнославянского происхождения; г) основы приставочных глаголов и т. п. Абсолютная нормализация предусматривает перевод пар регулярно варьирующихся букв (например, *о* / *ѡ*, *е* / *ѣ*) и буквосочетаний (например, *шт* / *щ*,

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

жю / жу) к единому представлению (например, *о*, *е*, *щ*, *жу*). При абсолютной нормализации унифицируются как единицы грамматического словаря, так и словоформы в тексте.

Введение

Данная статья является продолжением работы «К задаче автоматической лексико-грамматической разметки старорусского корпуса XV–XVII вв.»[1], опубликованной в одном из предыдущих выпусков «Вестника ПСТГУ». Старорусский корпус НКРЯ включает порядка 5000 документов, созданных преимущественно в XV–XVII вв., с некоторым добавлением текстов XIV в., а также текстов, широко датируемых кон. XVII в. — нач. XVIII в. Корпус включает как оригинальные тексты, так и воспроизведения документов раннего периода. В жанровом отношении это деловые документы, бытовая переписка, летописи, сказания, поучения, жития, Четьи-Минеи, другие памятники религиозной литературы и т. п.[2] В настоящее время онлайн-платформа корпуса (http://ruscorpora.ru/search-mid_rus.html) предоставляет возможность поиска по точной форме слова. Кроме того, можно ограничить поиск выборками текстов по жанру, времени создания и некоторым другим метатекстовым признакам. Планируется развитие функционала старорусского корпуса, с тем чтобы, во-первых, пользователь мог искать вхождения слов в упрощенной и модернизированной орфографии, а во-вторых, чтобы ему был доступен поиск по лемме (исходной форме слова) и грамматическим характеристикам словоформы. В связи с этим тексты корпуса будут размечены автоматически с помощью морфологического таггера и проиндексированы в поисковой системе корпуса.

Разметка словоформ осуществляется с привлечением грамматических словарей и баз данных. Вход грамматического словаря состоит из леммы и грамматического индекса или же из основы (набора основ) и грамматического индекса. Грамматические индексы находятся в особой зоне словаря и содержат информацию о части речи, постоянных (словоклассифицирующих) грамматических признаках, структуре парадигмы и окончаниях каждой грамматической формы. Таким образом, грамматический индекс позволяет сопоставить каждой лемме набор пар <словоформа — грамматическая характеристика>. В грамматической базе данных корпуса отражены непосредственные соответствия между словоформой и леммой, частью речи и грамматическими характеристиками.

В текстах Старорусского корпуса наблюдается значительная вариативность написаний словоформ. Так, одна и та же форма глагола может быть записана как *взіал, взялъ, възял, взялъ, взял, възялъ, взіалъ, взял* и *възялъ*. Нестабильная орфография памятников является причиной неэффективной работы автоматических морфологических анализаторов, в основе работы которых лежит принцип, что одной лексико-грамматической аннотации в норме должна соответствовать одна

[1] См.: *Гаврилова Т. С., Шалганова Т. А., Ляшевская О. Н.* К задаче автоматической лексико-грамматической разметки старорусского корпуса XV–XVII вв. // Вестник ПСТГУ. Сер. III: Филология. 2016. № 2. С. 7–25.

[2] См.: *Молдован А. М.* Памятники древнерусской письменности в Национальном корпусе русского языка // Труды Института русского языка РАН. 2015. Вып. 5. С. 88–98.

цепочка символов (за исключением отдельно оговариваемых случаев типа вариантов 3 л. ед. числа аориста, имперфекта и др.). Приведенный пример показывает, что в ситуации нестабильной орфографии одной лексико-грамматической основе соответствует множество орфовариантов основы и множество орфовариантов окончаний.

Таким образом, целью нашей работы было разработать модуль поддержки орфографической вариативности как дополнение к морфологическому таггеру.

Поддержка орфографической вариативности

Среди методов преодоления орфографической вариативности, применяемых в исторических корпусах, можно выделить два основных подхода: относительную и абсолютную нормализацию орфографии текста[3]. Относительная нормализация предполагает перечисление в словаре, к которому обращается анализатор, всех возможных вариантов написания словоформы, а абсолютная — нахождение наиболее похожей формы из всех, зафиксированных в словаре, через изменение расстояния Дамерау-Левенштейна[4], с применением спелл-чекера или каким-либо другим способом. Абсолютная нормализация орфографии, как правило, менее трудоемка в исполнении, однако плохо применима для текстов с сильной орфографической вариативностью. При наличии нескольких «ошибок» в одной словоформе абсолютная нормализация может привести к таким серьезным ошибкам, как изменение исходной леммы. К тому же такая модель нечувствительна к морфологическому членению слова и может исказить морфологический анализ, изменив флексию. В данной работе использованы оба способа (абсолютный и относительный) нормализации орфографии текстов.

Нормализация вариативности во флексиях

В связи с тем что орфографическая вариативность характеризует как основы, так и флексии, базу флексий было необходимо расширить. Например, именительный падеж слов многих парадигм оригинально оканчивался на *ъ*, однако часто можно встретить написание и без *ъ* (например, *раб* вместо *рабъ*). Таким образом, возникает необходимость добавления «нулевой флексии» в качестве алломорфа окончания *ъ* для данного разбора. Перечислим типы добавленных вариантов:

- добавление варианта флексии с *е* для всех флексий, у которых в оригинале был *ѣ* (в связи с историческим переходом *ѣ* в *е*);
- добавление вариантов *тьс, тц, ц* для всех флексий, содержащих сочетание *тс,* и вариантов *тс, тц, ц* для всех флексий, содержащих сочетание *тьс* (позволяет распознать такие глагольные формы как *делаеться,* однако может создавать дополнительную омонимию в случаях типа *молится,* ко-

[3] См.: *Piotrowski M.* Natural Language Processing for Historical Texts. Synthesis Lectures on Human Language Technologies. Vol. 17. San Rafael, 2012. P. 69–78.

[4] См.: *Jurafsky D.*, *Martin J. H.* Speech and language processing. International Edition. New Jersey, 2000. P. 107.

торое получит разбор не только как форма настоящего времени третьего лица (*мóлится*), но и как форма инфинитива (*молúться*));
- добавление вариантов без конечного ъ для всех флексий, оканчивающихся на ъ, в том числе добавление нулевых флексий в случаях, если ъ был единственным элементом окончания;
- добавление вариантов написания различных флексий с учетом передачи или не передачи йотирования гласных (например, *аа* вместо *ая — вернаа*).

Нормализация вариативности в именных основах определенных типов

Помимо случаев, когда одна и та же словоформа по тем или иным причинам имеет разную орфографическую запись, существуют случаи, когда одно слово могло иметь различные реализации в разные временные промежутки или в разных диалектах и, как следствие, записываться по-разному.

К таким случаям относятся появление особой основы во множественном числе прилагательных на *-ский* (например, *ростовстие* вместо *ростовские*), использование особых основ у прилагательных на *-еческий* (например, возможны формы *деческий* и *децкий* (также *детский* и *детцкий*)), использование различных форм существительных с сочетанием *жд* (например, *одежа* и *одежда*), использование различных основ у существительных на *-я* (*келия* и *келья*).

Строго говоря, подобные слова являются разными словоформами. Однако для удобства пользования корпусом в данной работе подобная вариативность считается вариативностью внутри одной лексемы.

Кроме того, написание или отсутствие ъ в позиции не перед гласным может являться как орфографической вариативностью (при отсутствии редуцированного звука, но выражении его на письме как дань письменной традиции), так и реальной вариативностью, не связанной с орфографией (с произнесением редуцированного или без произнесения). Таким образом, разграничение различных типов вариативности иногда является сложной задачей; в данной работе вне зависимости от причины возникновения различного написания, все слова, являющиеся вариантами реализации одной лексемы в определенной грамматической форме, считаются вариантами реализации одной словоформы.

Для слов вышеупомянутых типов автоматически были порождены и добавлены в словарь дополнительные основы. Например, слово *келия* получило в словаре две альтернативные основы — *кели* и *кель*.

При порождении основ были учтены некоторые регулярные расхождения между двумя типами написания основ, которые можно условно назвать «церковнославянским» и «древнерусским». Например, для слов с корнем *един* были порождены альтернативные основы с корнем *один*. Таким образом, были обработаны соответствия древнерусского *о* и церковнославянского *je* в начале слова (*один* / *един*, *озеро* / *езеро*), а также древнерусского *я* и старославянского *а* в той же позиции (*ягня* / *агнць*)[5].

[5] За основу был взят список соответствий. См.: *Винокур Т. Г.* Древнерусский язык. М., 1961.

Тем не менее одно из основных лексических отличий между церковнославянским и древнерусским — наличие полногласия — не было обработано. Причина заключается, прежде всего, в том, что далеко не все церковнославянские «краткие» варианты соответствуют «полным» русским. Например, цепочки *брат* и *хлѣб* никогда не имели полногласных аналогов, и автоматическое порождение полногласных основ привело бы к появлению в словаре таких лексем, как *борот* и *холоб*. Поэтому было принято решение не проводить автоматического порождения полногласных основ для церковнославянских вариантов вида СраС, СлаС, СрѣС, СлѣС.

Нормализация вариативности в глагольных основах

Нормализации подверглись приставочные глаголы. Приставки являются морфемами, в написании которых наблюдается особенно большая орфографическая вариативность. Даже в современном языке можно встретить варианты вроде *раздать* и *расдать*, хотя последний и не соответствует орфографической норме. Поэтому глаголам, обладающим определенным префиксом, были приписаны основы, содержащие альтернативные варианты написания этого префикса. Следующие варианты написания приставок занесены в грамматический словарь: *из — ис; раз — рас; воз — вз — вос — вс; рос — роз; бес — без; черес — чрес — чьрес — через — чрез — чьрез; с — з — со.*

Чтобы избежать порождения дополнительных основ для бесприставочных глаголов, начинающихся с буквы *с*, последнее преобразование проводилось только в случае, если за буквой следовал звонкий согласный.

Отглагольные существительные не обладают в грамматическом словаре специальным морфологическим тегом, и их невозможно определить автоматически. Поэтому, несмотря на то что в написании их приставок тоже присутствует вариативность, она не была обработана в текущей версии словаря.

Нормализация других случаев вариативности как вариантов основ в грамматическом словаре

Некоторые другие случаи орфографической вариативности также решались с помощью добавления дополнительных основ в словарь. К ним относятся:

- выпадение интервокального *г* (например, для слова *пáгуба* в словаре добавлена дополнительная основа *пауб*);
- отсутствие начального *г* в словах с корнем *господ* (например, для слова *господний* была добавлена дополнительная основа *осподн*);
- замена йотированной гласной на нейотированную пару при сочетании двух гласных (например, для слова *боярин* была добавлена дополнительная основа *боарин*);
- ненаписание двойного согласного в корне (например, для слова *русский* была добавлена дополнительная основа *руск*);
- вариативность окончания наречий на *енне-онно* (например, для слова *безбоя́зненнѣ* была добавлена дополнительная основа *безбоязненно*);

- выпадение согласных в консонантном кластере: выпадение *т* в сочетаниях *стн, зтн, нтс*; *д* в сочетании *сдн, здн*; *г* в сочетании *гск*; *к* в сочетании *кск* (например, для слова *праздный* была добавлена дополнительная основа *празн*). Это упрощение может приводить к уменьшению точности: например, слово *косный* может быть проанализировано как форма прилагательного *костный*;
- написание консонантных кластеров разными способами. Для слов, содержащих сочетание *жч* и *сч* был добавлен вариант с *щ*. Например, для слова *счастье* в словаре добавлена основа *щасть*.

Нормализация орфографии с помощью общих правил

Различия орфографической системы, не покрытые добавлением новых основ в грамматический словарь, были решены с помощью абсолютной нормализации орфографии текста. К словоформам корпуса в указанном порядке применялись следующие правила:

- *і*, *ї* (*і* кириллическая) и *и* соответствует буква *и*;
- *о* и *ѡ* соответствует буква о;
- *у* и *ѹ* (ук) соответствует буква *у*;
- *я*, *ѧ* и *іа* соответствует буква *я*;
- *ф* и *ѳ* соответствует буква *ф*;
- *е* и *ѣ* в основах соответствует буква *е;*
- *уо* и *оу* соответствует буква *у*;
- *жы*, *шы, щы* и *чы* соответствуют сочетания *жи*, *ши*, *щи* и *чи,* соответственно;
- *жю*, *щю* и *шю* соответствуют сочетания *жу*, *щу* и *шу,* соответственно;
- *шт* соответствует *щ*;
- *дч* и *дщ* соответствует *дш*;
- после всех согласных, не занимающих позицию перед гласной, ставится *ъ,* за исключением согласных, занимающих позицию конца слова.

Уточним, что эти правила были применены как к словоформам в тексте, так и к основам в грамматическом словаре.

После применения указанных правил всем встречающимся в корпусе словоформам, соответствующим современному *взял* (*взіал, взѧлъ, възял, взялъ, взѧл, възялъ, взіалъ, взял* и *възѧлъ*), будет соответствовать одна форма *възял*.

Заключение: технологическая цепочка разметки

Автоматическая разметка текстов среднерусского корпуса будет проходить в четыре этапа. Сначала будут опознаны словоформы, присутствующие в древнерусском корпусе НКРЯ[6], и их лексико-грамматические характеристики (все зафиксированные сочетания <часть речи — лемма — набор грамматических помет>) будут перенесены в старорусский корпус. Затем будут опознаны «со-

[6] В древнерусском корпусе НКРЯ реализована ручная морфологическая разметка. См.: *Мишина Е. И., Пичхадзе А. А.* Древнерусский подкорпус Национального корпуса русского языка // Труды Института русского языка РАН. 2015. Вып. 6. С. 99–115.

временные» словоформы, а именно те, которые покрываются автоматическим разметчиком Mystem, основанным на грамматическом словаре А. А. Зализняка[7]. На третьем этапе словоформы, оставшиеся неразмеченными, будут проанализированы с помощью программы Юни-таггер Т. А. Архангельского[8]. Программа работает на основе специально разрабатываемого грамматического словаря для письменности старорусского периода. Наконец, словоформы, которые остались вне покрытия старорусского грамматического словаря, будут проанализированы с помощью статистического разметчика Tree Tagger[9] (возможно, с последующей ручной коррекцией лемм). Как показали предварительные эксперименты, такой гибридный подход обеспечивает максимальную полноту и более высокую точность разметки.

Однако обратим внимание на то, что каждая из описанных составляющих разбора имеет свои преимущества и недостатки, причем на каждом следующем шаге полнота возрастает, но точность падает. Например, словоформа, совпадающая со словоформой древнерусского корпуса, иногда может иметь в тексте другую грамматическую интерпретацию. Часть парадигм, распознаваемых анализатором современного русского языка, может пересекаться с парадигмами других лексем, присутствующих в текстах XV–XVII вв., но отсутствующих в современном грамматическом словаре. Заметим, правда, что хотя такие ложные срабатывания возможны, но тем не менее доля их в общем корпусе форм ничтожна. Разметка на основе грамматического словаря старорусской письменности достаточно адекватно покрывает формы словоизменения частотной лексики нашего корпуса, однако вследствие особенностей технологии словоформе приписываются как актуальные, так и «паразитические» разборы (например, для несуществующих лемм, построенных автоматическим способом)[10]. Статистический разметчик, строящий гипотезы на основе вероятности последовательных цепочек частеречных тегов в тексте и соответствий между окончанием словоформы и окончанием леммы, обеспечивает 100 % полноты разборов, но характеризуется наименьшей точностью.

Согласно данным, приведенным в нашей работе[11], мы ориентируемся на точность распознавания части речи порядка 89–94 % и точность распознавания

[7] См.: *Зализняк А. А.* Грамматический словарь русского языка: Словоизменение. М.: Русский язык, 1977. 4-е изд., испр. и доп., М.: Русские словари, 2003; *Segalovich I.* A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. Proceedings of MLMTA, Las Vegas, Nevada, 2003. P. 273–280. В данном случае Mystem работает без модуля порождения гипотез для несловарных слов.

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

[9] См.: *Schmid H.* Probabilistic part-of-speech tagging using decision trees. Proceedings of the International Conference on New Methods in Language Processing, 1994.

[10] При составлении грамматического словаря изначально не ставилось ограничений на порождение таких лишних разборов. Во-первых, предполагается, что пользователь вряд ли будет искать несуществующие леммы в корпусе. Во-вторых, в дальнейшем планируется создать модуль, уменьшающий неоднозначность разборов с учетом вероятности комбинаций тегов в контексте.

[11] См.: *Гаврилова Т. С., Шалганова Т. А., Ляшевская О. Н.* К задаче автоматической лексико-грамматической разметки старорусского корпуса XV–XVII вв. С. 7–25. См. также:

леммы порядка 75–79 %. Мы планируем провести экспертизу качества лексико-грамматической разметки на текстах разных жанров и разного времени создания. Особое внимание будет уделено точности определения грамматических признаков (падежа, рода, числа и т. п.). Это та зона, в которой наблюдается существенное отставание компьютерно-лингвистических технологий.

Ключевые слова: древнерусский язык, старорусская письменность, корпус, НКРЯ, лексико-грамматическая разметка, орфографическая вариативность, орфоварианты, нормализация орфографии.

Список литературы

Архангельский Т. А. Принципы построения морфологического парсера для разноструктурных языков: Дис. ... канд. филол. наук. М.: МГУ, 2012.

Винокур Т. Г. Древнерусский язык. М.: Высшая школа, 1961.

Гаврилова Т. С., Шалганова Т. А., Ляшевская О. Н. К задаче автоматической лексико-грамматической разметки старорусского корпуса XV–XVII вв. // Вестник ПСТГУ. Сер. III: Филология. 2016. № 2. С. 7–25.

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

Мишина Е. И., Пичхадзе А. А. Древнерусский подкорпус Национального корпуса русского языка // Труды Института русского языка РАН. 2015. Вып. 6. С. 99–115.

Молдован А. М. Памятники древнерусской письменности в Национальном корпусе русского языка // Труды Института русского языка РАН. 2015. Вып. 6. С. 88–98.

Berdichevskis 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).

Jurafsky D., Martin J. H. Speech and language processing. International Edition. New Jersey, 2000.

Piotrowski M. Natural Language Processing for Historical Texts. Synthesis Lectures on Human Language Technologies. Vol. 17. San Rafael, 2012. P. 69–78.

Schmid H. Probabilistic part-of-speech tagging using decision trees. Proceedings of the International Conference on New Methods in Language Processing. 1994.

Segalovich I. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. Proceedings of MLMTA, Las Vegas, Nevada, 2003. P. 273–280.

Berdichevskis 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). М., 2016.

St. Tikhon's University Review.
Series III: Philology.
2017. Vol. 51. P. 11–20

Gavrilova Tatiana,
National Research University "Higher School of Economics"
20 Miasnitskaia str., Moscow, 101000, Russian Federation
tanya96gavrilova@yandex.ru

Shalganova Tatiana,
National Research University "Higher School of Economics"
20 Miasnitskaia str., Moscow, 101000, Russian Federation
insana7@mail.ru

Liashevskaia Ol'ga,
Candidate of Science in Philology,
National Research University "Higher School of Economics",
Institute of Russian Language, Russian Academy of Science
20 Miasnitskaia str., Moscow, 101000, Russian Federation
olesar@yandex.ru

Взіалъ, възялъ, вьзял: Processing Orthographic Variation in Lexico-Grammatical Annotation of the Middle Russian Corpus of 15th–17th Centuries

T. Gavrilova, T. Shalganova, O. Liashevskaia

This paper discusses the problem of heterogenous orthography in Middle Russian texts in terms of their automatic processing. The Middle Russian subcorpus of the Russian National Corpus contains documents written mainly between 1400 and 1700, when spelling variation was still wide-spread. The task of lexico-grammatical analysis is to assign a dictionary form (lemma), a part of speech indication and grammatical tags to each word form in the corpus. Traditional methods of grammatical tagging depend on the fact that there usually only one string of characters that represents the stem and the ending of each grammatical word form. Because of this, heterogenous orthography leads to errors in the work of automatic morphology analysers (taggers) if they are not provided with the module that supports orthographic variation.

In this project, both relative and absolute normalisation is used. Relative normalisation involves multiplying orthographic representations of stems and endings in the grammatical dictionary according to standard rules. This is carried out on the level of (a) word endings; (b) nominative stems with regular variation, e.g. *russk(ij) / russt(ij), keli(ja) / kel'(ja)*; (c) nominative stems of Church Slavonic origin, e.g. *odin-* / *edin-*; (d) verb stems with prefixes, etc. Absolute normalisation matches characters (or character combinations) that alternate regularly in the corpus (e.g. *о/ѡ* 'omega', *е/ѣ, шт/щ, жю/жу*). Absolute normalisation is applied both to orthographic representations in the grammatical dictionary and to word forms in the text.

Keywords: Middle Russian, Old Russian, Russian National Corpus, lexico-grammatical tagging, morphological analysis, spelling variation, unstable orthography, orthographic normalisation.

References

Berdichevskis A., Eckhoff H. M., Gavrilova T. 2016. The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian, in: *Komp'iuternaia lingvistika i intellektual'nye tekhnologii*, 15 (22).

Gavrilova T. S., Shalganova T. A., Lyashevskaya O. N., K zadache avtomaticheskoi leksiko-grammaticheskoi razmetki staro-russkogo korpusa XV–XVII vv., in: *Vestnik PSTGU, Series III: Philology*, 2016, Vol. 47 (2), 7–25.

Jurafsky D., Martin J. H., *Speech and language processing*. International Edition. New Jersey.

Mishina E. I., Pichkhadze A. A., Drevnerusskiĭ podkorpus Natsional'nogo korpusa russkogo iazyka, in: *Trudy Instituta russkogo iazyka RAN*, 2015, 6, 99–115.

Moldovan A. M., Pamiatniki drevnerusskoĭ pis'mennosti v Natsional'nom korpuse russkogo iazyka, in: *Trudy Instituta russkogo iazyka RAN*, Moscow, 2015, 6, 88–98.

Piotrowski M., *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Vol. 17. San Rafael, CA, 69–78.

Schmid H., Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*.

Segalovich I., A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. *Proceedings of MLMTA, Las Vegas, Nevada*, 273–280.

Vinokur T. G., *Drevnerusskii iazyk*, *Old Russian Language*, Moscow, 1961.

Zalizniak A. A., *Grammaticheskii slovar' russkogo iazyka: Slovoizmenenie Grammatical Dictionary of the Russian Language: Inflection*, Moscow, 1977. 4th edition: Moscow, 2003.