FILOLOGIJA 68, Zagreb 2017.

UDK 81'374=161.1
https://dx.doi.org/10.21857/ygjwrcj56y
Izvorni znanstveni članak
Primljen 30.IV.2017.
Prihvaćen za tisak 23.X.2017.

Анна Пичхадзе
Институт русского языка им. В. В. Виноградова
Отдел лингвистического источниковедения
и истории русского литературного языка
ул. Волхонка 18/2, RU-1190119 Москва, Россия
rusyaz@yandex.ru

РАЗМЕТКА ЦЕРКОВНОСЛАВЯНСКИХ И ДРЕВНЕРУССКИХ ТЕКСТОВ: ПРОБЛЕМЫ ЛЕММАТИЗАЦИИ

В статье описываются проблемы лемматизации, возникающие при электронной разметке древнерусских текстов, и способы оформления проблемных лемм: введение дополнительных полей, альтернативный разбор и объединение разных вариантов леммы в одной словарной статье. Каждый из способов является оптимальным в разных ситуациях. Для различения семантических омонимов и видовых пар глаголов достаточно введения дополнительных полей. Если отсутствуют критерии, на основании которых можно было бы однозначно реконструировать лемму из-за орфографических или фонетико-орфографических эффектов, свойственных древним памятникам письменности, целесообразно применять альтернативный разбор, допускающий восстановление нескольких лемм для одной словоформы. В случаях, когда варьированию подвержена только словарная форма, предлагается использовать специальный алгоритм лемматизации: занесение в словарь леммы в её исконном (древнейшем) виде вместе со всеми позднейшими вариантами, причем все позднейшие варианты указываются также в специальном поле и автоматически связываются с исходной леммой. Такой алгоритм обеспечивает переадресацию к древнейшему виду леммы, даже если при разметке будет выбран позднейший вариант.

При электронной разметке средневековых текстов возникают разнородные проблемы. Одни из них обусловлены экстралингвистическими причинами: они связаны с недостаточностью наших сведений

о языке древности и с тем, что аннотируемые памятники могут не содержать форм, необходимых для однозначного восстановления леммы. Другие проблемы связаны с вариативностью лемм: словарная форма может изменяться со временем или варьировать в зависимости от диалекта. Со всеми этими проблемами, однако, сталкиваются и составители традиционных (бумажных) словарей и словоуказателей. Но если для традиционных словарей выработаны способы подачи лемм, которые не восстанавливаются однозначно, то для электронных словарей вопрос оформления таких лемм пока остается достаточно острым. Практика электронной разметки древних славянских текстов сравнительно молода, и способы оформления неоднозначно восстанавливаемых лемм пока еще только опробуются. Кроме того, при электронной разметке нестандартная лемматизация требует разработки специальных алгоритмов, обеспечивающих её корректность. Ниже описываются подходы к лемматизации, применяемые в Институте русского языка Российской академии наук (Москва) при разметке древнерусских текстов.

Семантическая омонимия

Легче всего решается при лемматизации проблема семантической омонимии. Если омонимы относятся к разным частям речи (например, существительное *лъжь* ‘ложь’ и прилагательное *лъжь* ‘лживый’), их разведение происходит автоматически, поскольку частеречная характеристика является обязательной в любой системе разметки: лемматизация возможна только при указании части речи размечаемого слова. Если же омонимы принадлежат к одной и той же части речи, для их дифференциации достаточно ввести какое-нибудь дополнительное поле с указанием признака, который их различает. Если слова различаются семантикой, это будет поле «Значение»: с его помощью будут разведены, например, леммы *оутрьнии* ‘утренний’ и *оутрьнии* ‘внутренний’, *градъ* ‘город’ и *градъ* ‘осадки’, *състоупитисѧ* ‘сойтись, столкнуться’ и *състоупитисѧ* ‘оступиться, поскользнуться’ и т. п.

Однако проблема с различением семантических омонимов полностью решается при помощи дополнительного поля только при разметке одного отдельного памятника. При сведении размеченных текстов в корпуса возникают новые трудности. Дело в том, что поле «Значение» факультативно, оно заполняется не при всех леммах, а только при необходимости различить омонимы. Но, размечая древ-

ние тексты, мы не знаем всех омонимов, которые существовали в языке и которые могут встретиться в размечаемом тексте. Между тем неожиданное появление нового омонима в ходе разметки — явление нередкое: например, наряду с широко распространенным *помощи* ‘помочь’ в древнерусском переводе «Истории Иудейской войны» Иосифа Флавия встретился глагол *помощи* ‘смочь’ (ИИВ II 2004:250), не зафиксированный словарями, а в древнерусском переводе «Жития Андрея Юродивого» наряду с общеизвестным *надѣꙗтисѧ* ‘надеяться’ — гапакс *надѣꙗтисѧ* ‘налегать, оказывать давление, ἐπανάκειμαι’ (ЖАЮ:431, строка 5740). Историческими словарями русского языка хорошо засвидетельствован глагол *съзирати* ‘обозревать, осматривать’ (СРЯ 2002:106), но в древнерусской «Пчеле» встретился омонимичный глагол, который больше нигде пока не зафиксирован: *съзирати* ‘созревать (о зерне)’ (Пчела II 2008:320). Это означает, что, во-первых, при глаголах *помощи* ‘смочь’, *надѣꙗтисѧ* ‘налегать’ и *съзирати* ‘созревать’ обязательно должно быть заполнено поле «Значение». Во-вторых, поле «Значение» должно быть заполнено и для глаголов *помощи* ‘помочь’, *надѣꙗтисѧ* ‘надеяться’ и *съзирати* ‘обозревать, осматривать’, хотя до появления новых омонимов заполнять его не было необходимости. При этом, если размечаемый текст входит в корпус из разных текстов, поле «Значение» для глаголов *помощи* ‘помочь’, *надѣꙗтисѧ* ‘надеяться’ и *съзирати* ‘обозревать, осматривать’ должно быть заполнено во всех текстах данного корпуса. Таким образом, появление каждого нового омонима требует редактирования разметки всех размеченных в корпусе текстов и их общего словаря, что в свою очередь ставит разработчиков перед необходимостью создавать специальные программы для синхронизации словарей разных памятников. По мере того, как пополняется их общий словарь, в каждом отдельном тексте у всё большего числа слов заполняется поле «Значение» за счёт разметки семантических омонимов.

Видовые пары глаголов

Определенные проблемы возникают при лемматизации омонимичных глаголов, выражающих разные видовые значения. Если глаголы с разным видовым значением имеют одинаковый инфинитив при разном презенсе, их легко различить, введя дополнительное поле «Форма настоящего времени»: тогда глагол *засыпати* с презенсом *засыплютъ* можно отделить от глагола *засыпати* с презенсом *засыпаютъ*, глагол *написати* с презенсом *напишоутъ* от глагола *написати*

с презенсом *написаютъ*, глагол *наскакати* с презенсом *наскачоутъ* от глагола *наскакати* с презенсом *наскакаютъ* и т. п. Однако не все глаголы с разным видовым значением и одинаковым инфинитивом различаются формой презенса. Например, инфинитив *напитати* соотносится с единственной формой презенса *напитаютъ*, которая может иметь два разных видовых значения: 'накормят' и 'кормят'. Для разведения таких омонимов можно использовать дополнительное поле «Вид» (совершенный или несовершенный): *напитати*, *напитаютъ* (совершенный вид) и *напитати*, *напитаютъ* (несовершенный вид).

Впрочем, такой подход не бесспорен. Категория вида в средневековый период находится в развитии, и простых критериев отнесения глагола к совершенному или несовершенному виду не существует. Применительно к древнему периоду славянской письменности вопрос о видовой принадлежности должен решаться индивидуально для каждого глагола. В частности, исторические словари трактуют глагол *напитати* во всех значениях — 'кормить' и 'накормить' — как одну лексему, считая глагол двувидовым. Однако в древнерусский период процесс становления категории вида зашел уже достаточно далеко, так что в ряде случаев допустима и альтернативная трактовка — выделение двух омонимичных глаголов. Заполнение поля «Вид» в ходе разметки, помимо технической роли разведения омонимов, может выполнять и другую полезную функцию — накопления информации о видовых глагольных парах и, тем самым, о развитии категории глагольного вида.

Альтернативные леммы

В предыдущих разделах речь шла о проблемах, связанных с оформлением леммы, но не с её восстановлением. Между тем главные трудности при разборе древних текстов появляются, когда отсутствуют критерии, на основании которых можно было бы однозначно реконструировать лемму. Такая ситуация часто возникает как результат орфографических или фонетико-орфографических эффектов, свойственных древним памятникам письменности. Случаи, когда на основании того или иного написания невозможно восстановить единственно возможную словарную форму, подразделяются на две группы: к первой относятся слова, различающиеся во всех своих формах, ко второй — слова, различающиеся только в некоторых формах.

Слова, различающиеся во всех формах

В эту группу входят прежде всего слова, часто встречающиеся в сокращенных написаниях. Если в памятнике фиксируются формы *благодѣть* и *благодать* и наряду с ними — сокращенные формы с выносной буквой «д» типа *блг^{д}ть*, то для сокращенных форм одинаково возможна лемматизация *благодѣть* или *благодать*. Бывает, что у возвратных глаголов в размечаемом тексте встречаются только формы с конечным выносным «с» под титлом (типа *съжалити^{с}*). Поскольку известно, что в древнерусском глаголы «горестного чувства» и некоторые другие глаголы образовывали возвратные формы как с аффиксом *-си*, так и с аффиксом *-сѧ* (Крысько 1995:478—480), для форм типа *съжалити^{с}* правомерна как реконструкция *съжалитиси*, так и реконструкция *съжалитисѧ*.

Особые трудности вызывают буквенные сокращения. Обычно буквами под титлом обозначаются числительные, и контекст, в котором встречается буква в значении цифры, не всегда позволяет определить, имеется ли в виду количественное или порядковое числительное. Однако буква под титлом может обозначать и наречие. Например, буква «г» под титлом, помимо числительного, нередко употребляется также как обозначение наречия со значением ‘трижды, три раза’. Поскольку существуют многочисленные словообразовательные варианты наречий с указанным значением (*тришьды, тришьда, троичи, троицею*), каждый из них теоретически может претендовать на роль леммы для сокращенного буквенного написания. Чтобы решить, какая именно лемма скрывается за буквенным написанием, нужно иметь перед глазами словесные написания данного наречия в размечаемом тексте; однако такая информация не всегда доступна, к тому же в памятнике могут употребляться две синонимичные леммы одновременно, — в результате буквенное написание не поддаётся однозначной лемматизации.

Много неоднозначных форм возникало в древних рукописях в результате падения и прояснения редуцированных. Одним из следствий прояснения редуцированных было устранение древних чередований в корне. Например, после прояснения *ъ → о* старая форма прилагательного *довъльныи* совпала с формой *довольныи*, которая содержала гласный в другой ступени чередования. Поэтому в рукописях, возникших после прояснения редуцированных, форма *довольныи* двусмысленна: она может отражать исконное *довольныи* и исконное *довъльныи* с проясненным редуцированным в корне. Такие же трудно-

сти возникают с глаголами *тъснитисѧ* ‘тщиться, стараться, стремиться’ и *тьснитисѧ* ‘то же’, различающимися вокализмом корня: в рукописях, созданных после падения редуцированных, встречаются написания типа *тнѧхоусѧ* (имперфект 3 лица мн. числа), которые могут быть возведены как к исходной форме *тъснитисѧ*, так и к *тьснитисѧ* (ИИВ II 2004:420). Еще сложнее интерпретировать производные с приставкой *об-*, этимологически безъеровой, но еще до падения и прояснения редуцированных развившей по аналогии с другими приставками вторичный *-ъ-* на конце, поскольку в древности у неё существовал и вариант *обь-* (откуда с удлинением гласного возникал вариант *оби-*). Если в рукописи, отражающей падение редуцированных, одновременно встречаются три варианта — *обходити*, *объходити* и *обьходити*, написания типа *обходити* могут восходить к трём леммам: древней безъеровой форме *обходити*, к форме с вторичным редуцированным *объходити* и к форме с исконным редуцированным *обьходити*.

В период второго южнославянского влияния в восточнославянских рукописях сочетания редуцированных с плавными стали писаться с редуцированным после плавного, причем качество редуцированного не различалось. Поскольку многие древнерусские памятники дошли до нас в рукописях XV-XVI вв., в них представлены написания типа *тръгноути* ‘дёрнуть, рвануть’ или *кръстица* ‘коробочка, шкатулка’, которые можно интерпретировать либо как *търгноути* и *кърстица*, либо как *тьргноути* и *кьрстица*, потому что и тот, и другой корень имел варианты с корневым *-ъ-* и *-ь-*.

Аналогичные трудности возникают в восточнославянских памятниках из-за смешения *е* и *ѣ* и (позднее) *о* и *а*. Смешение *е* и *ѣ* носило как чисто графический, так и фонетический характер; независимо от своего характера, оно приводило к неразличению, например, корневого гласного у глаголов *метати* (презенс *мечоуть* и *метаютъ*) ‘бросать, кидать’ и *мѣтати* (презенс *мѣтаютъ*) ‘то же’. Соответственно в рукописях со смешением *е* и *ѣ* формы *метаютъ* и *мѣтаютъ* могут предполагать как инфинитив *метати*, так и инфинитив *мѣтати*. В рукописях, отражающих аканье, возникают такие же проблемы при лемматизации форм с чередованием: например, формы имперфекта типа *престоѩше, престоѩхоу* могут соотноситься как с инфинитивом *престоѩти*, так и с инфинитивом *престаѩти*.

В традиционных словарях и указателях в перечисленных случаях используются разные способы оформления заголовочной статьи: иногда в заголовок выносится несколько лемм («**метати** или **мѣтати**», «**тъснитисѧ** и **тьснитисѧ**»), иногда оформляется несколько статей, связанных

взаимными отсылками («**престоѭати**» смотри/сравни «**престаѭати**»). При электронной разметке эти способы неприменимы. Однако программные средства позволяют сделать альтернативный разбор любой словоформы, причем количество вариантов разбора не ограничено:

Словоформа: *довольныи*	
Вариант разбора 1	Вариант разбора 2
Лемма: **довъльныи**	Лемма: **довольныи**
Род: мужской	Род: мужской
Число: единственное	Число: единственное
Падеж: именительный	Падеж: именительный

При поиске словоформа, связанная с двумя (или несколькими) леммами одновременно, будет найдена независимо от того, какая из альтернативных лемм выбрана для поиска.

Слова, различающиеся в некоторых формах

Нередко приходится сталкиваться с ситуацией, когда слово употреблено в такой форме, которая может быть возведена к разным леммам, причём в памятнике не встречаются формы, позволяющие сделать однозначный выбор, или же встречаются альтернативные леммы одновременно. У прилагательных с суффиксами *—ьныи* и *-ьнии* позицией нейтрализации различий является именительный падеж множественного числа: форма им. мн. *вьрховьнии* допускает реконструкцию исходной формы в виде *вьрховьнии* или *вьрховьныи*. Для приставок *об-* и *оби-* позицией нейтрализации является положение перед *и-*: два *-ии-* подвергались стяжению, поэтому презенс *обидоуть* ‘обходят, окружают’ может соотноситься с инфинитивом *обити* или *обиити*, в то время как в положении перед согласным приставки различаются, ср. причастия прошедшего времени *ошьдъ* и *обишьдъ*.

Есть случаи, когда вариативность восходит к праславянской эпохе. Так, инфинитив *ристати* характерен для южнославянских памятников, *рискати* — для восточнославянских. Каждая из двух форм имеет свои параллели в балтийских языках (Фасмер III:485—486, 530) и в силу этого претендует на роль самостоятельной леммы, но в презенсе (*рищоуть* и т. п.) их различить невозможно.

Другие варианты возникали в силу диахронических изменений позже, в исторический период. Глагол *жрѣти*, *жрьжтъ* ‘приносить жертву’ уже в Супрасльской рукописи обнаруживает признаки перехода в другой тип спряжения — *жрѣти*, *жьрѫтъ* (Вайан 1952:312—

313). В древнерусских памятниках тоже встречаются формы как исконного, так и вторичного типа: *жьрти*, *жьрють* и *жрети*, *жьроуть*. Однако в одной из самых частотных форм — 3 лица ед. числа — оба типа склонения имеют одинаковую форму *жьреть*, поскольку мягкость *-р-*, как правило, не обозначается. Соответственно форма может быть отнесена как к исконному, так и вторичному типу.

Двусмысленными бывают некоторые формы слов, относящихся к разным типам склонения: например, форма винительного падежа множественного числа *доумьцѧ* или *доумьцѣ* или местного падежа единственного числа *доумьци* может восходить либо к *доумьца* 'советник', либо к *доумьць* 'то же'. Слова могут иметь разный род, как в случае существительных *глезнъ* 'лодыжка' (м. род), *глезна* 'то же' (ж. род) и *глезно* 'то же' (ср. род), — формы мужского и среднего рода различаются только в прямых падежах, для женского рода показательных форм гораздо больше. У некоторых экзотизмов невозможно определить род даже по исходной форме: было ли географическое название *Дафнии* мужского или женского рода, нельзя установить не только по форме местного падежа, но и по форме именительного падежа, потому что греческие имена на -η передавались в славянских языках существительными женского рода на *-и* (они склонялись как *мълнии*), которые переосмыслялись как имена мужского рода по аналогии с существительными типа *жрѣбии*.

Иногда ни одна падежная форма не позволяет установить род существительного: и леммы, и словоформы омонимичны. Так, *гъртань* в древности изменялось по *i*-склонению как мужского, так и женского рода. Единственным показателем рода в данном случае является определение (*въ моемь гъртани*), и, если оно отсутствует в тексте, формы могут быть отнесены к лемме как мужского, так и женского рода.

Во всех перечисленных случаях при электронной разметке целесообразно использовать альтернативный разбор, т. е. связать двусмысленную словоформу с альтернативными леммами:

Словоформа: *гъртани*	
Вариант разбора 1	Вариант разбора 2
Лемма: **гъртань**, Род: мужской	Лемма: **гъртань**, Род: женский
Число: единственное	Число: единственное
Падеж: местный	Падеж: местный

Варианты леммы

У многих слов словарная форма по разным причинам подвержена вариациям — в то время как все остальные формы вполне стабильны. Так, у личных имен мужского рода *о*-склонения именительный падеж мог оканчиваться на *-ъ* и на *-о*; обе формы отмечены в старославянских и древнерусских памятниках: *Марко* и *Маркъ* (СС 1994:323; ИИВ II 2004:115). В таких случаях использовать альтернативный разбор было бы неправомерно: здесь нет лемм, различающихся словарными характеристиками (родом существительных), словообразовательными элементами (разными приставками или суффиксами) или огласовкой морфем, проходящей по всей парадигме слова; здесь вариативность затрагивает только словарную форму. Альтернативный разбор есть смысл использовать, когда мы имеем дело с разными словами, здесь же более уместно говорить о вариантах леммы.

Локальные варианты

Варианты лемм часто обусловлены отличиями между церковнославянскими и восточнославянскими рефлексами тех или иных сочетаний звуков, реже — диалектными различиями внутри восточнославянского ареала. Поскольку рефлекс праславянского сочетания **kti* в южнославянских и восточнославянских языках был разным, инфинитивы корневых глаголов на заднеязычный согласный в древнерусских памятниках принимают разный вид в зависимости от ориентации на церковнославянскую норму или восточнославянский узус: *лещи* или *лечи* при презенсе *лѧгоуть*, *рещи* или *речи* при презенсе *рекоуть*, *тещи* или *течи* при презенсе *текоуть* и т. п. Аналогичным образом варьируют инфинитивы с полногласием /неполногласием: *братисѧ* или *боротисѧ* при презенсе *борютьсѧ* — с той только разницей, что здесь возможен ещё и восточнославянский диалектный (северо-западный) инфинитив *бротисѧ*.

В отличие от южнославянских, восточнославянские инфинитивы обычно не отражают эффект третьей палатализации. Поэтому презенсу *движоуть* может соответствовать инфинитив *двигати* или *двигати* и т. п. Южнославянским языкам свойственно сохранение губных согласных перед глагольным суффиксом **nǫ*, а восточнославянским — их утрата, отсюда расхождение между (*по*)*гыбноути* и (*по*)*гыноути*, которое сохраняется в презенсе (*погыбноуть*/*погыноуть*), но нейтрализуется в прошедших временах (*погыбоша*, *погыбъшии* и др.).

Диахронические варианты

Со временем лемма во многих случаях подвергалась аналогическим изменениям: так, исконная форма именительного падежа единственного числа существительного *цьркы* ‘церковь’ в древнерусских текстах иногда принимает вид *цьркъви* по аналогии с косвенными падежами. В славянской традиции словарной формой глагола является инфинитив, но именно инфинитив, во-первых, особенно часто подвергается аналогическим перестройкам, а во-вторых, не является частотной формой, так что для некоторых глаголов инфинитивы вообще не засвидетельствованы. Лишь гипотетически восстанавливается нигде не зафиксированный инфинитив (*по*)*шити* ‘(по)бить’ для форм типа (*по*)*шибоуть*, (*по*)*шибоша*. По поводу того, как мог бы выглядеть инфинитив, соответствующий презенсу *скорбоу* (3 лицо мн. числа) ‘скорбят, сокрушаются’, встретившемуся в одной из берестяных грамот, не один год ведутся дискуссии (Зализняк 2004:448; Крысько 2016:19).

Древние чередования, противопоставлявшие основу инфинитива основе презенса, устранялись уже в старославянских канонических текстах: здесь отмечен выровненный по аналогии с презенсом инфинитив *писати* наряду с *пьсати*, переход от первоначального инфинитива *льзати* к *лизати* (Вайан 1952:301). В церковнославянских текстах число перестроенных форм возрастает. В древнерусских памятниках наряду с исконным инфинитивом *чисти* ‘читать’ употребляется выровненный по настоящему времени инфинитив *чьсти* (*чести*), поэтому формы типа *чьтоуть*, *чьтошѧ* и т. д. могут быть связаны как с леммой *чисти*, так и леммой *чьсти*. То же относится к презенсу типа *смѣютьсѧ*: он может соотноситься с исконным инфинитивом *смиѧтисѧ* или с перестроенным по аналогии с презенсом *смѣѧтисѧ*. Поскольку списки, в которых дошли до нас древние памятники, часто существенно отстоят от оригинала во времени, в рукописях встречаются — наряду с исконными — еще более поздние образования: *съмотрѣти* наряду с *съмотрити*, *плыти* наряду с *плоути*, *(по)грести* и даже (*по*)*гребсти* наряду с (*по*)*грети*.

Перестройки касаются не только глаголов с чередованием: аналогическое взаимодействие иногда происходит и между приставками, в результате чего, например, наряду с исконным *выити* ‘выйти’ по аналогии с *сънити* ‘сойти’ появляется *вынити*, так что формы причастий прошедшего времени типа *вышедъ* могут быть равным образом отнесены к инфинитиву *выити* или *вынити*.

Во всех перечисленных случаях нестабильности леммы, какими

бы причинами ни была вызвана вариативность, приходится использовать специальный алгоритм лемматизации. Он заключается в том, что в словарь заносится лемма в исконном виде вместе со всеми позднейшими вариантами, причем все позднейшие варианты указываются также в специальном поле и автоматически связываются с исходной леммой:

Лемма в словаре	Варианты леммы
грети / грести / гребсти	грести, гребсти
чисти / чьсти	чьсти

Такой алгоритм обеспечивает переадресацию к древнейшему виду леммы, даже если при разметке будет выбран позднейший вариант, и представляет собой аналог отсылки в традиционном (бумажном) словаре: *чьсти* смотри *чисти*.

***

Итак, при разметке средневековых славянских текстов для решения проблем лемматизации можно использовать разные способы: введение дополнительных полей, альтернативный разбор и приравнивание вариантов леммы. Разумеется, исследователь сам решает, какой способ оптимален в каждом конкретном случае. Иногда их приходится комбинировать: если дополнительные поля не обеспечивают корректную разметку, осуществляется еще и альтернативный разбор. Альтернативный разбор — выход из любой сомнительной ситуации, в том числе и той, когда неясно, к какому омониму относится та или иная словоформа, то есть когда сомнительна семантика слова или его частеречная принадлежность. Распространение практики электронной разметки наверняка приведет к совершенствованию способов решения проблем лемматизации.

Литература

Вайан, Андре. 1952. *Руководство по старославянскому языку*. Москва: Издательство иностранной литературы. 446 str.

Зализняк, Андрей А. 2004. *Древненовгородский диалект*. Москва: Языки славянской культуры. 867 str.

ИИВ I-II 2004. *«История Иудейской войны» Иосифа Флавия: Древнерусский перевод*. Изд. подгот. А. А. Пичхадзе, И. И. Макеева, Г. С. Баранкова, А. А. Уткин. Т. I—II. Москва: Языки славянской культуры. 879 str.; 837 str.

Крысько, Вадим Б. 1995. Залоговые отношения. U: *Древнерусская грамматика XII—XIII вв*. Москва : «Наука». 465—506.

Крысько, Вадим Б. 2016. Чередование *о/ø* в глагольном формообразовании. U: *Грамматические процессы и системы в синхронии и диахронии. Тезисы докладов международной конференции* (*30 мая — 1 июня 2016 г.*). Москва. 18—19. (http://ruslang.ru/doc/thesis.pdf)

Молдован, Александр М. 2000. *«Житие Андрея Юродивого» в славянской письменности*. Москва: «Азбуковник». 759 str.

Пчела I—II 2008. *«Пчела»: Древнерусский перевод*. Изд. подгот. А. А. Пичхадзе, И. И. Макеева. Т. I-II. Москва: Рукописные памятники Древней Руси. LXVI, 444, 304 str.; 634 str.

СРЯ 2002. *Словарь русского языка XI—XVII вв*. Вып. 26. Москва: «Наука». 277 str.

СС 1994. *Старославянский словарь* (*по рукописям X—XI веков*). Под ред. Р. М. Цейтлин, Р. Вечерки и Э. Благовой. Москва: «Русский язык». 841 str.

Фасмер, Макс. 1987. *Этимологический словарь русского языка*. Т. III. Москва: «Прогресс». 831 str.

Obilježavanje crkvenoslavenskih i staroruskih tekstova — problemi lematizacije

Sažetak

Članak opisuje probleme lematizacije koji nastaju prilikom računalnoga obilježavanja tekstova na staroruskom jeziku i načine sređivanja problematičnih lema kao što su unos dodatnih polja, alternativno prepoznavanje, ujedinjavanje svih oblika određene leme u istom rječničkom članku. Kako bi se razlikovali semantički homonimi, odnosno parovi glagola ovisno o vidu, dovoljno je uvesti dodatna polja. Ako ne postoje kriteriji na temelju kojih je moguće na jedinstveni način rekonstruirati lemu — zbog pravopisnih, odnosno izgovorno-pravopisnih čimbenika prisutnih u starim pismenim spomenicima —, poželjno je rabiti alternativnu morfološku analizu koja dopušta uspostavljanje nekoliko lema za isti oblik riječi. Kada se varira samo oblik riječi, predlaže se koristiti se posebnim algoritmom lematizacije, a to je unošenje u rječnik leme u njezinu izvornom (najstarijem) obliku ukupno sa svim njezinim mlađim varijantama. Pri tome se svi mlađi oblici navode u posebnom polju i automatski se povezuju s izvornom lemom. Taj algoritam pruža preusmjeravanje prema najstarijem obliku leme, čak i u slučaju odabiranja najmlađe varijante prilikom obilježavanja.

Ключевые слова: Грамматическая разметка текстов, лемматизация, церковнославянский язык, древнерусский язык

Ključne riječi: Računalno obilježavanje, lematizacija, crkvenoslavenski jezik, staroruski jezik[1]

[1] Статья подготовлена в рамках работы над проектом РГНФ № 14—04—12035в «Система электронной грамматической разметки древнерусских и церковнославянских текстов».