Баранов Виктор Аркадьевич
Ижевский государственный технический университет
имени М.Т.Калашникова
УДК 519.2:801.82(045)

Из опыта создания электронного грамматического словаря древнерусского языка: модель, реализация, использование

древнерусский язык, исторический корпус, лемматизация

Электронный грамматический словарь исторического корпуса «Манускрипт» (manuscripts.ru) предназначен для обеспечения работы автоматического лемматизатора славянских текстов X–XVII веков.

Модель базы данных словаря позволяет хранить информацию о леммах древнерусского языка, состоящих из двух компонентов – изменяемой и неизменяемой частях словоформ (граница между ними может не совпадать с лингвистической границей), а также о связях между ними. (Псевдо)окончания (псевдо)основ, имеющих идентичные (под)парадигмы, объединены в уникальные группы (типы изменения – ТИ), содержащие сведения о морфологических характеристиках, общих для всех форм (под)парадигмы. Основа леммы может иметь подчиненные основы с собственными типами изменения и наборами окончаний.

Приведем фрагмент парадигмы леммы *ходити*:

ходи – ТИ глаг. – *ти* инф. / *тъ* суп.

хож – ТИ глаг., изъяв., наст. – *оу* ед., 1 л.

хожд – ТИ глаг., изъяв., наст. – *оу* ед., 1 л.

ход – ТИ глаг., изъяв., наст. – *иши* ед., 2 л. / *ить* ед., 3 л. / —

ходи – ТИ глаг., изъяв., аор. – *хъ* ед., 1 л. / *Ø* ед., 2 л. / *Ø* ед., 3 л. / *хомъ* мн., 1 л. / *сте* мн., 2 л., / *ша* мн., 3 л. / —

хожа – ТИ глаг., изъяв., имперф. – *ахъ* ед., 1 л. / *хъ* ед., 1 л. / *аше* ед., 2 л. / *ше* ед., 2 л. / —

ходѧ – ТИ глаг., изъяв., имперф. – *ахъ* ед., 1 л. / *хъ* ед., 1 л. / *аше* ед., 2 л. / *ше* ед., 2 л. / —

хожда – ТИ глаг., изъяв., имперф. – *ахъ* ед., 1 л. / *хъ* ед., 1 л. / *аше* ед., 2 л. / *ше* ед., 2 л. / —

ходи – ТИ глаг., повел. – *Ø* ед., 2 л. / *Ø* ед., 3 л. / *мъ* 1 мн., 1 л. / *те* мн., 2 л. / *вѣ* дв. 1 л. / *та* дв., 2 л.

ход – ТИ прич., наст., действ., имен. – *ѧ* муж., ед., им.

ходѧч – ТИ прич., наст., действ., имен. – *ь* муж., ед., вин. / *а* муж., ед., род. / *а* муж., ед., род.-вин. / —

ходѧщ – ТИ прич., наст., действ., имен. – *ь* муж., ед., вин. / *а* муж., ед., род. / *а* муж., ед., род.-вин. / — и т.д.

Леммам, имеющим одинаковое количество основ и идентичные наборы их окончаний, присваивается один и тот же индекс; подчиненные

основы могут иметь дополнительные индексы. Так, лемма *ходити*, а также еще 135 лемм с основой на *д* с идентичной парадигмой имеют индекс *а1_д*, а подчиненные основы их причастий – *а1_прич1_д*, *а1_прич2_д* и др. (см. также [Баранов 2007; Баранов 2010]).

Реализация парадигмы в виде перечня главной и проиндексированных подчиненных основ и связанных с ними групп окончаний позволяет пополнять и редактировать словарь, а в случае необходимости визуализировать отдельные подпарадигмы леммы, например, подпарадигмы нечленных (именных) и членных (местоименных) форм, подпарадигмы причастий, а также изменять состав итогового словника текста или подкорпуса, давая причастия отдельными статьями.

Процедура поиска леммы для текстовой формы и присвоения последней морфологических помет осуществляется с помощью отдельного модуля корпуса – автоматического морфологического анализатора – и заключается в сопоставлении текстовой формы с единицами словаря. Нормализованная подача форм парадигмы в словаре и максимально точная транскрипция текстов корпуса потребовала разработки приемов нахождения соответствующих по грамматическому значению, но различных по написанию слов.

Основным приемом является применение и к единицам словаря, и текстовым прецедентам единых правил унификации, приведение тех и других к условным формам и поиск для текстовых форм полных соответствий в словаре. Правила включают устранение диакритики и титл, приведение заглавных и выносных букв к строчным, унификацию а) отдельных букв – *о* = *ѡ*, *и* = *і*, *ф* = *ѳ* и др., б) йотированных и нейотированных букв в позициях начала слова и после гласных – *ю* = *оу*, *ꙗ* = *а* и др., в) буквенных сочетаний – *ша* = *шꙗ* = *шѧ* = *шіа*, *жоу* = *жю* = *жѫ* = *жѭ*, *че* = *чѥ* и др., г) лигатур, диграфов и монографов – *ѿ* = *от*, *оу* = *у* и др., д) буквенных групп – *согласный* + *р*/*л* + *ѣ* + *согласный* = *согласный* + *р*/*л* + *е* + *согласный* и другие приравнивания. Всего используется более 20 правил (см. их перечень на странице http://manuscripts.ru/mns/slov.list_preobr).

Другим способом является добавление в базу данных словаря форм, отличия которых от нормализованных форм парадигмы не описаны правилами: *здоровьѥ* – *съдоровиѥ*, *хитрость* – *хытрость* и под., – а также форм под титлом: тр͡ца, п͡ъ, ст҃го и под. При включении таких форм в словарь может быть указано: использовать форму или для определенного текстового прецедента, или для конкретной морфологической формы, или для всех омонимичных форм парадигмы.

Третий прием предполагает хранение в словаре вариантных лемм одного (заимствованного) слова – *моисии* – *моисеи* – *моисѣи* – *моиси*, *симеонъ* – *сьмеонъ* – *семеонъ* – *сумеонъ* – *семьюнъ* – *симонъ* и под., установление в базе данных связи между ними, приведение текстовых прецедентов к соответствующим начальным формам и демонстрацию

всех вариантов в одной словарной статье или формирование разных с перекрестными отсылками. Этот способ установления соответствия между текстовым прецедентом и единицей словаря используется также и для незаимствованных слов, например, с начальными *преже-* / *пережде-* / *прежде-* / *прѣже-* / *прѣжде-*, с различными суффиксами – *обитель* – *обитѣль*, с различными основами – *законодавьць* – *законоудавьць* и др.

В лемматизаторе реализовано несколько режимов приведения к леммам текстовых форм и определения их морфологических признаков: автоматический анализ (автоматическая лемматизация), автоматическое снятие омонимии, ручное снятие омонимии. При первом лемматизатор соотносит текстовую форму со всеми соответствующими ей омонимичными формами словаря, включая в итоговый перечень соответствий внутри- и межлеммные. Второй – обеспечивает приведение текстового прецедента к одной и только одной лемме и предлагает наиболее вероятную форму парадигмы [Баранов 2020]. Третий – позволяет пользователю вручную выбрать (подтвердить) вариант разбора текстовой формы.

До недавнего времени модули демонстрации корпуса (однотекстовая и многотекстовая запросные формы, статистические модули) обеспечивали визуализацию словников только с использованием результатов лемматизации без снятия омонимии. В таких словниках с помощью условных обозначений и перекрестных отсылок указывается на неоднозначность определения морфологических признаков текстовых прецедентов. Приведем пример трех статей из Толстовского сборника XIII в. (РНБ, F.п.I.39), лл. 100–184:

адамовъ {**137.2-17 см. адамъ}

адамъ (8) {**106.2-14}, {**126.1-8}, {**137.2-17 см. адамовъ}, {**150.1-1 см. адъ}, {**159.1-4}, {**164.2-21 см. адъ}, {**167.2-5}, {**180.1-1}

адъ (11) {**107.2-7}, {**126.1-25}, {**131.2-4}, {**131.2-5}, {**137.1-7}, {**147.2-15}, {**150.1-1 см. адамъ}, {**164.2-21 см. адамъ}, {**171.2-7}, {**173.1-16}, {**173.1-19},

где { } – адрес текстовой формы с неснятой омонимией, ** – несколько разборов текстовой формы в пределах одной парадигмы, см. – несколько разборов в пределах парадигм двух или более лемм.

Использование режима автоматического снятия омонимии позволяет с высокой степенью точности устранить повторение одной текстовой формы в разных статьях:

адамъ (8) {106.2-14}, {126.1-8}, {137.2-17}, {150.1-1}, {159.1-4}, {164.2-21}, {167.2-5}, {180.1-1}

адъ (9) {107.2-7}, {126.1-25}, {131.2-4}, {131.2-5}, {137.1-7}, {147.2-15}, {171.2-7}, {173.1-16}, {173.1-19}.

Работа над словником текстов Кирилла Туровского из Толстовского сборника потребовала не только снятия омонимии вручную, но и добав-

ления в словарную базу данных большого количества форм под титлом и лемм, отсутствующих в словаре.

Приведем пример статьи леммы *хотѣти* в ее сегодняшнем виде (лл. 1–48):

хотѣти (28) глаг (парадигма a2a_т*)

изъяв. наст. ед. 1 л. (6) хощю 7.1-17, 9.2-12, 10.1-19, 12.1-25, 19.2-16, 19.2-18, *2 л.* хощеши 17.2-12, *3 л.* (2) хощеть 10.2-3, 10.2-13, *прич. наст. действ. ед. муж. им.* хотѧ 3.2-12, *жен. род.* хотѧща 20.1-24, *прош. ед. муж. им.* хотѣлъ 17.2-13, *жен. им.* хотѣла 7.1-3

& хощю *изъяв. наст. ед. 1 л.* 24.2-16, хощеши *2 л.* 48.2-6, хощеть *3 л.* (2) 27.2-15, 47.1-1, хощемъ *мн. 1 л.* 32.1-21, хощете *2 л.* | *дв. 3 л.* (4) 24.1.1-6.22, 29.1.1-23.2, 29.2.1-2.22, 29.2.1-5.7, хотѧ *прич. наст. действ. ед. муж. им.* (3) 24.1-8, 24.1-19, 46.1-16, хотѧщаго *род.-вин.* | *род.* 31.1-14, хотѧ *средн. им.* (3) 24.1-8, 24.1-19, 46.1-16, хотѧще *вин.* 36.1-1, хотѧщаго *род.-вин.* | *род.* 31.1-14, хотѧще *мн. муж. им.* 36.1-1, хотѣли *прош. мн. муж. им.* | *жен. им.* 24.1.1-11.7,

где & – начало раздела с неснятой вручную омонимией, | – варианты разбора.

Реализованная в системе «Манускрипт» модель словарной базы данных, ее первичное наполнение, возможности редактирования и пополнения, режимы модуля морфологического анализа и форм демонстрации результатов автоматической и ручной лемматизации позволяют осуществить весь цикл работ по созданию словника текстов корпуса «Манускрипт».

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (РФФИ) в рамках проекта «Подготовка интернет-издания и комплексное исследование языка и письма Толстовского сборника XIII в. (РНБ, F.п.I.39)» (проект № 18-012-00428).

Литература

Баранов В.А. Автоматический морфологический анализатор древнерусского языка: лингвистические и технологические решения / В.А.Баранов, А.Н.Миронов, А.Н.Лапин, И.С.Мельникова, А.А.Соколова, Е.А.Корепанова // 10-я юбилейная международная конференция «EVA 2007 Москва». – Москва, 2007. – URL: http://conf.evarussia.ru/upload/eva2007/reports/ doklad_1318.pdf.

Баранов В.А. Морфологическая парадигма и ее составляющие в системе «Манускрипт» / В.А.Баранов, О.В.Гулина, А.Н.Миронов // Информационные технологии и письменное наследие: материалы междунар. науч. конф. (Уфа, 28–31 октября 2010 г.) / отв. ред. В.А.Баранов. – Уфа; Ижевск: Вагант, 2010. – С.28–31. – URL: https://textualheritage.org/english/el-manusctipt-10/18.html.

Баранов В.А. Инструменты извлечения и приемы подготовки лингвистических данных для статистического анализа в историческом корпусе «Манускрипт» / В.А.Баранов, Р.М.Гнутиков // Труды XVII Международной конференции Ассоциации «История и компьютер» (в печати)