Благодарности

Работа выполнена при финансовой поддержке РФФИ (проект № 11–01–00251, № 12–01–00481, № 12–07–00070), поддержке РГНФ (проект № 12–04–12062), проекта № 213 Программы фундаментальных исследований Президиума РАН «Интеллектуальные информационные технологии, математическое моделирование, системный анализ и автоматизация» и проекта № 2.2 Программы ОНИТ РАН «Интеллектуальные информационные технологии, системный анализ и автоматизация». Спасибо Николаю Тесля за плодотворное обсуждение экспериментальной части работы.

Список литературы

Гришина и др., 2005 — Гришина, Е. А., Плунгян В. А. Перспективы развития Национального корпуса русского языка // Национальный корпус русского языка. М.: Индрик, 2005. http://www.ruscorpora.ru/corpora-biblio.html.

Крижановский, 2010 — Крижановский А.А. Преобразование структуры словарной статьи Викисловаря в таблицы и отношения реляционной базы данных // Препринт. 2010. http://scipeople.com/publication/100231/.

Крижановский, 2011а — Крижановский А.А. Количественный анализ лексики английского языка в викисловарях и Wordnet // Труды СПИИРАН. 2011. Вып. 19. С. 87–101. http://scipeople.com/publication/106012/

Крижановский, 2011б — Крижановский А.А. Оценка использования корпусов и электронных библиотек в Русском Викисловаре // Труды международной конференции «Корпусная лингвистика–2011». СПб., 2011. С. 217–222. http://scipeople.com/publication/102432/.

Леденёва, 2008 — Леденёва В.В. Лексикография современного русского языка. Практикум: Учеб. пособие. М., 2008.

Meyer and Gurevych, 2012 — Meyer C.M., Gurevych I.. Wiktionary: a new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography // Electronic Lexicography. Oxford: Oxford University Press. 2012. (в печати). http://www.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2011/oup-elex2012-meyer-wiktionary.pdf.

МОРФОЛОГИЧЕСКАЯ МОДЕЛЬ КАЛМЫЦКОГО ЯЗЫКА В СВЕТЕ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ: ПРЕДВАРИТЕЛЬНЫЕ ЗАМЕЧАНИЯ

В. В. Куканова

Калмыцкий институт гуманитарных исследований РАН, Элиста

The article is devoted to the description of morphological model of the Kalmyk language in the light of automatic processing of texts — for the creation of morphological parser and lemmatizator. The author considers the problems of revealing of morphological paradigms for productive work of parser.

Автоматическая обработка текстов невозможна без работы морфологического анализатора, которая может осуществляться только на основе лингвистической информации. Создание анализатора основывается или на словарном, или на бессловарном подходе. Как нам кажется, для калмыцкого языка необходимо развивать эти два подхода параллельно, поскольку тот грамматический словарь, который сейчас уже подготовлен, основан на словнике калмыцко-русского словаря [КРС, 1977]. В нем отсутствуют многие единицы, составляющие ядро лексической системы языка. Бессловарный модуль требуется и для повышения количественных данных разборов, он «может предсказать морфологические характеристики практически любого слова, если парадигма попадает под одну из хранимых» [Автоматическая, 2011: 118].

Но здесь и кроется вся сложность автоматической обработки текстов, язык которых принадлежит к монгольской группе: практически невозможно выделить точные парадигмы словоизменения, поскольку язык является агглютинативным по своей структуре. Во-первых, главная особенность подобных языков заключается в теоретической возможности присоединения в строгом порядке неограниченного количества словоизменительных аффиксов к основе слова. Во-вторых, определенный тип парадигмы выделяют на основе общих грамматических категорий и словоизменительных аффиксов, а в русском языке еще и частей речи (к примеру, субстантивное, адъективное и местоименное склонения). В калмыцком языке имеет место четкое противопоставление именного и глагольного словоизменения: для каждого из них существуют неомонимичные аффиксы, не считая частиц, которые могут присоединяться к любой части речи. Для всех именных частей речи — существительных, местоимений (за исключением некоторых форм с супплетивными основами), числительных — используются те же самые словоизменительные аффиксы для выражения категорий числа, падежа и др. По сути, можно выделить единое склонение для именных частей речи. В этом случае уместно вспомнить об алгоритмах работы морфологического анализатора, который может анализировать слова слева направо (от начала к концу слова) и справа налево (от конца слова к началу). Для агглютинативного языка более всего подходит первый способ анализа, так как основа в калмыцком языке в основном неизменяема, за исключением ряда слов. К тому же изменения в основе достаточно предсказуемы, например, если слова заканчиваются на -н, то происходит его усечение при соединении определенного аффикса. Словоизменительные аффиксы же могут совпадать со словообразовательными, что часто бывает в калмыцком языке, а это, как известно, ведет к неправильным разборам. В-третьих, морфотактические правила одинаковы для сочетания как основы со словоизменительным аффиксом, так и словоизменительного аффикса с другим аффиксом, т. е. при словоизменении (как склонении, так и спряжении) на морфемных швах происходят одни и те же процессы.

Противопоставление неизменяемых и изменяемых классов слов также не продуктивно, поскольку почти все они по своей форме являются изменяемыми: в калмыцком языке почти все части речи имеют потенциальную

возможность изменения и перехода из одной части речи в другую, следовательно, последний процесс коренным образом влияет на грамматические и дистрибутивные характеристики слов в контекстах (ср.: «части речи не могут быть строго делимы на изменяемые и неизменяемые…» [Котвич, 1929: 86]). Примечательно, что еще в начале XX в. В.Л.Котвич отметил, что слов действительно «неизменяемых» немного в калмыцком языке (см.: «Вообще совершено неизменяемых слов в калмыцком языке имеется очень мало: если в каком-либо положении слово не изменятся, то в другом положении оно может принять приставку соответствующую склонению или спряжению, и таким образом сделаться изменяемым. Неизменяемыми остаются только служебные частицы и междометия» [Котвич, 1929: 86]). В этом аспекте уместно вспомнить прилагательные и причастия, которые легко субстантивируются. Что касается наречий, послелогов, то они могут присоединять к себе личнопритяжательные частицы, или аффиксы поссевности. Например, на ‘на этой стороне, на эту сторону’ и на-нь ‘на этой стороне; еще, больше’, деер ‘на, возле, около; пока; вместе с, с’ и деернь ‘пока’. Поэтому некоторые части речи можно назвать условно неизменяемыми. По нашим наблюдениям, к прилагательным могут присоединяться аффиксы сказуемости, а к наречиям и послелогам — аффиксы поссесивности. К собственно неизменяемым принадлежат частицы, междометия и звукоподражания (идеофоны).

Несмотря на все это, мы, тем не менее, попытались установить словоизменительные парадигмы в калмыцком языке. Критериями для выделения типов именного словоизменения послужили: количество основ, морфологические процессы на стыке основы и словоизменительного аффикса, происхождение слов и его слоговая структура, аффикс множественного числа (сочетание семантического фактора с фонетическим), сингармонизм. Что касается парадигм глагольного словоизменения, то здесь учитывалось следующее: количество основ, морфонологические процессы на стыке основы и словоизменительного аффикса, происхождение слова и его слоговая структура, сингармонизм.

Модель словоизменительных классов калмыцкого языка создавалась на базе обратного словаря, позволившего достаточно быстро определить частеречную принадлежность слов, например, слова, которые заканчиваются на -х, являются в основной своей массе глаголами. К тому же рядом стоящие лексические единицы имеют сходную парадигму словоизменения, на стыке основы и словоизменительных аффиксов в словах происходят те же самые морфонологические процессы [Белоногов, 1967; Зализняк, 1987: 9].

Первоначально словник включал чуть более 25 тыс. входов, т. е. вокабул, но в ходе анализа словник значительно уменьшился, поскольку были извлечены все словоформы, например Genitive, Ablativ и другие падежные формы, а также атрибутивные формы глагола (причастия и деепричастия). Были выделены и соответственно сформированы лексиконы следующих частей речи

(приводятся также их традиционные термины[1]): ADJ (Имя прилагательное/Adjective), ADV (Наречие/Adverb), CONJ (Союз/Conjunction), INJ (Междометие/Interjection), N (Имя существительное/Noun), NUM (Числительное/Numeral), PART (Частица/Particle), POST (Послелог/Postposition), PRON (Местоимение/Pronoun), V (Глагол/Verb) и U (Неизвестная грамматическая категория/Unknown category). Группа неизменяемых слов (наречия, послелоги, союзы, междометия, частицы) далее не подвергалась анализу, поскольку не могут изменяться по тем или иным грамматическим категориям. Однако следует отметить возможность присоединения к ним частиц разной функциональной нагруженности (аффиксы сказуемости, или личнопредикативные частицы, вопросительные, модальные и др.).

Ряд лексем был продублирован, поскольку часть из них имеет различную частеречную принадлежность, и по своей сути они являются грамматическими омонимами: в отличие от русского языка, в котором происходят совпадения в рамках одной словарной статьи, в калмыцком языке имеет место совпадение форм разных грамматических классов. В словаре они получили дополнительные индексы, сигнализирующие о том, что данная единица может иметь несколько вариантов разборов в зависимости от контекста (та же самая операция проводилась при обработке лексических омонимов). Например, модн ‘дерево’ и модн ‘деревянный’. Для их разграничения (снятия омонимии) можно использовать дистрибутивный метод, т. е. учитывать окружение анализируемой единицы: если справа стоит существительное, то скорее всего анализируемая единица является прилагательным.

Таким образом, морфологическая система калмыцкого языка опирается на несколько взаимозависимых и взаимообусловленных компонентов:

1) лексиконы, в котором даются лемма и ее возможные стеммы (графические основы слова), например, слово ханлт ‘благодарность; удовлетворение’ может быть представлена одной основой ханлт-, то ‘количество; цифра; учет; номер’ двумя стеммами то- и тоо-, дун ‘голос, песня, звук’ — дуу-, дуун-, дун-;

2) таблица словоизменительных аффиксов (окончания приводятся в графической форме), а также частиц, которые могут примыкать к слову, например, -нр — аффикс множественного числа (Pl), -ан — возвратная частица (PART.REFL), -шц — частица уподобления (PART.EQU);

3) таблица словоизменительных моделей по имени существительному и глаголу[2], например, N1 — основа + аффикс множественного числа (эмч+нр ‘врачи’); N2 — основа + аффикс падежа (эмч+үр ‘к врачу’); N12 — основа + аффикс множественного числа + аффикс падежа (эмч+нр+ин ‘врачей’);

4) таблица происхождения слов, где на основе помет S1 (исконно калмыцкие или заимствованные, но фонетически адаптированные) и S2 (заимст-

[1] Только в этом случае корпус будет доступен не только специалистам в области лингвистики, но и преподавателям калмыцкого языка, школьникам, что необходимо в возрождении главных функций языка в обществе – когнитивной и коммуникативной.

[2] Требуется для создания генератора, основной целью которого является проверка правильности разборов анализатора.

вованные в последние десятилетия и фонетически не адаптированные, имеющие только одну специфическую черту в словоизменении, которая заключается в том, что сингармонизм гласных осуществляется не по первому гласному, а по последнему, который определяет качество последующих гласных звуков (букв) в слове) выделяются мягкий и твердый варианты словоизменения.

Список литературы

Автоматическая, 2011 — Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др.. М. 2011.

Белоногов и Давыдова, 1967 — Белоногов Г.Г., Давыдова И.М. О возможности определения грамматических классов слов по буквенным кодам слов // НТИ. Сер. 2. 1967. № 8.

Зализняк, 1987 — Зализняк А.А. Предисловие // Зализняк А.А. Грамматический словарь русского языка: Словоизменение: Около 100 000 слов. 3-е изд., стереотип. М.: Изд-во «Русский язык», 1987. С. 9.

КРС, 1977 — Калмыцко-русский словарь / Под ред. Б. Д. Муниева. М., 1977.

Котвич, 1929 — Котвич В.Л . Опыт грамматики калмыцкого разговорного языка. Изд. 2-ое. Ржевнице у Праги, 1929. 418 с.

К ПРОБЛЕМЕ ЭЛЕКТРОННЫХ СЛОВАРЕЙ КАЛМЫЦКОГО ЯЗЫКА

В. В. Куканова, Е. В. Бембеев

Калмыцкий институт гуманитарных исследований РАН, Элиста

The article is devoted to the problems of electronic dictionaries of Kalmyk language which belongs to the group of the endangered languages. The authors consider that it is necessary to create and develop the Kalmyk-Kalmyk dictionary with sound module and elements of thesaurus for the frequent words.

Сегодня пристальное внимание лингвистов всего мира направлено на использование информационных технологий в своих исследованиях. Не исключение в этом ряду и калмыцкое языкознание, для которого теория и практика составления электронных словарей является одной из актуальнейших проблем, поскольку в Республике Калмыкия сложилась ситуация утраты языка как средства познания и коммуникации. Калмыцкий язык причисляют к группе языков, которые находятся на грани исчезновения, поэтому существует настоятельная необходимость создания всевозможных электронных словарей, в первую очередь калмыцко-калмыцких, а также переводных (калмыцко-русских и русско-калмыцких).