В. В. Куканова

РАЗРЕШЕНИЕ ЛЕКСИКО-ГРАММАТИЧЕСКОЙ НЕОДНОЗНАЧНОСТИ В НАЦИОНАЛЬНОМ КОРПУСЕ КАЛМЫЦКОГО ЯЗЫКА: ПРЕДВАРИТЕЛЬНЫЕ ЗАМЕЧАНИЯ*

Введение

В связи с бурным развитием компьютерной лингвистики в отечественной науке результаты корпусных разработок стали активно применяться и для национальных языков народов Российской Федерации. Идея создания Национального корпуса калмыцкого языка (НККЯ) не могла не появиться и в калмыцком языкознании: в конце 2010 г. сотрудники КИГИ РАН приступили к ее осуществлению. Институт обратился с просьбой в редакцию газеты «Хальмг үнн» и издательский дом «Герел», которые любезно предоставили свои электронные архивы, ставшие заделом для корпуса «первого» порядка, т. е. коллекции текстов на калмыцком языке без морфологического аннотирования. Сейчас корпус начитывает почти 17 млн токенов: оцифрован значительный объем текстов на калмыцком языке, чего ранее никогда не предпринималось в калмыцком языкознании в целях получения материала для исследования с применением компьютерных технологий. Созданы программа, которая автоматически производит графематический, морфологический и семантический анализ, корпусный менеджер (система запросов и

* Статья подготовлена при финансовой поддержке РГНФ в рамках проекта «Национальный корпус калмыцкого языка: создание и разработка» (12-04-12047/в).

хранения информации, обработки данных и выдачи результатов). Это все было осуществлено за сравнительно небольшой срок (чуть более чем за четыре года)[1]. Перспективы развития корпуса калмыцкого языка видятся прежде всего не только в том, чтобы создавать новые подкорпусы и новые типы аннотаций, но и в том, чтобы снять омонимию, что существенно повысит качество разработанного информационного ресурса.

Морфологический анализатор современного калмыцкого языка, который создавался в ходе выполнения темы НИР, дает множественные вероятностные разборы, как и, впрочем, ожидалось: любой парсер, на каком бы языке программирования он не был написан, и независимо от того, для какого языка он предназначался. Этот факт повлиял на качество автоматически проведенного грамматического анализа, следует отметить, что количество множественных разборов весьма значительно в аспекте создания структурно-вероятностной модели калмыцкого языка, но не значительно в общем объеме текстов, поскольку это является традиционной проблемой для анализаторов, проводящих автоматический разбор текста. Парсер разбирает однозначно около 70–75 % токенов и дает множественный результат анализа у 15–20 % текстового материала. Эти данные, как можно заметить, варьируются в пределах 5 % по двум причинам: во-первых, при загрузке различных текстов в парсер их лексическое наполнение всегда не совпадает; во-вторых, словарь аффиксов не является полным в силу того, что были выделены наиболее продуктивные модели словоизменения, что определялось по обратному списку словоформ, непродуктивные остались вне поля зрения.

Множественные разборы, появляющиеся в ходе автоматического анализа линейной последовательности, и являются объектом внимания в данной работе. Здесь мы не касаемся таких теоретических вопросов, как подходы и способы разграничения омонимов, причин появления омонимии, различия омонимии и полисемии и др. Под омонимами понимается явление совпадения звуковой и/или письменной форм речи, обладающее разными лексическими и/

[1] Сравним с Национальным корпусом русского языка, который также был создан за этот же срок, при этом у разработчиков уже имелся большой объем оцифрованных текстов, грамматический словарь, отсутствовало лишь программное обеспечение.

или грамматическими значениями. Следует оговориться, что омонимы в калмыцком языке могут иметь не только сходную звуковую оболочку, но и одинаковое написание, хотя могут различаться по произношению. Это обусловлено тем, что, как замечают В. И. Рассадин и С. М. Трофимова, письменная и устная формы калмыцкого языка совершенно разные [2013], в основу письма не был положен фонемный принцип, что породило определенные проблемы. Из состава буквенных символов абсолютно несправедливо исключены так называемые редуцированные звуки, или «неясные гласные», хотя их фонемный статус был доказан экспериментально-инструментальными методами [Биткеев 1975].

Наличие омонимов в языке порождает появление многозначности автоматически проведенного морфологического анализа. Проблема разрешения многозначности является важным шагом в создании и разработке информационно-справочной системы, автоматической обработки текстов. Калмыцкий язык принадлежит к агглютинативным, то есть к слову для выражения грамматических значений в определенном порядке присоединяются аффиксы, выражающие одно грамматическое значение. Данная особенность влияет на алгоритм работы морфологического парсера (см. подробнее [Куканова, Каджиев 2014]).

Осуществить выбор между гипотетическими разборами, полученными в ходе автоматического анализа, вручную практически невозможно, поскольку объем созданного корпуса достаточно большой и обрабатывать данное количество целесообразнее автоматически, то есть при помощи модуля снятия омонимии, который «…может повысить точность обработки некоторых классов запросов и/или сократить объем хранимой информации» [Зеленков и др. 2005]. Для реализации этих целей необходимо разработать алгоритм снятия омонимии для калмыцкого языка.

Омонимия как таковая характерна для многих языков и имеет несколько видов, однако для письменного текста (линейной последовательности буквенных и цифровых символов) на калмыцком языке актуальны следующие типы омонимов[1].

[2] В статье не рассматриваются омофоны, так как их появление характерно для устной речи, в случае с письменным текстом они не носят релевантного характера.

1. Лексические омонимы. Например, слова *он*[1] ‘год; годовщина’, *он*[2] ‘насечка, зарубка’; *ааг*[1] ‘язычок; зазубрина остроги’, *ааг*[2] ‘высокомерие, гордость’, *ааг*[3] ‘развилка’, *ааг*[4] ‘настой’[1]. Их относят к полным омонимам: принадлежат одной части речи и имеют одинаковые парадигмы словоизменения.

2. Морфологические омонимы. К примеру, токен *давад* может относиться к лемме *дава*[1, 2, 3, 4, 5] (N) или *давх* (V). В первом случае токен будет иметь граммему N.Dat, а во втором — Conv.Ant. В лексикографии морфологические омонимы традиционно называют омоформами — словами, которые совпадают в одной грамматической форме и произносятся одинаково. Однако, что касается калмыцкого языка, то здесь они пишутся одинаково, но произносятся по-разному (ср.: *давад* [davādə] и [davād]). В большой степени они похожи на омографы.

3. Частеречные омонимы (образованные в ходе конверсии). Наиболее частотны такие омонимы среди прилагательных и существительных. Например, *модн* ‘дерево; деревянный’ в зависимости от окружения может являться существительным (N) или прилагательным (ADJ).

Таким образом, проблема для текстов на калмыцком языке заключается в снятии не только частеречной омонимии[2], но и лексической и морфологической омонимии, имеющей достаточно большие объемы в калмыцком языке, как и в случае с русским языком.

История вопроса

Существуют два подхода в снятии омонимии: детерминированные и вероятностные. Первый подход основан на правилах, которые выделены в ходе анализа корпуса текстов и введены в программный код. Второй подход базируется на применении вероятностных моделей («системы, построенные на основе статистических методов, снимают омонимию в текстах на этапе морфологического анализа, используя статистику совместной встречаемости грамматических признаков слов…» [Порохнин 2013]). Например, применение

[3] Здесь и далее переводы лексем приводятся по Калмыцко-русскому словарю [КРС 1977].

[4] Ср. с английским языком, где также изобилуют изафетные конструкции.

скрытой модели Маркова, методика опорных векторов, на основе нормализующих подстановок и позиций соседних слов. Выделяют также и третий подход, который синтезирует два вышеуказанных способа, так называемый гибридный. Но при этом каждый из них тестируется на корпусе текстов с уже снятой вручную омонимией.

Для агглютинативных языков снятие омонимии — проблема новая и неразработанная, поскольку сравнительно недавно созданы корпуса текстов, разработаны морфологические парсеры (корпус татарского, башкирского, шорского, монгольского, бурятского и других языков). Только сейчас перед разработчиками встает проблема снятия омонимии, после того, как был создан парсер, мы можем уже проанализировать результаты работы морфологического парсера и сформировать словарь омонимов. Практика разработки модулей по снятию омонимии в отечественной науке в основном касалась только русского языка, который, напомним, принадлежит к группе флективных. В мировой науке несомненно более разработанными являются модули для английского языка, испанского и др., т. е. тех языков, которые носят международный характер. Отметим, что разрешению лексико-грамматической неоднозначности уделяли внимание и в турецком языкознании. Турецкий язык является наиболее близким в структурном отношении калмыцкому языку, поэтому часть алгоритмов и подходов, видимо, можно применить и к калмыцкому языку.

Предварительные замечания

К сожалению, словарь омонимов калмыцкого языка до сих пор не разработан, что несколько усложняет работу по снятию омонимии. Лексические омонимы были извлечены из Калмыцко-русского словаря под ред. Б.Д. Муниева [1977], единственной лексикографической работы академического характера. В словаре основ каждый лексический омоним записывается отдельно, анализатор же приписывает все возможные разборы токену. В случае со снятием лексической омонимии каким-то образом сложно спрогнозировать конкретное значение в каждом отдельном случае. Возможно, необходимо задействовать в работе контекстный подход: учесть левое и правое окружения токена, проанализировать дистрибуцию с точки

зрения семантики и сочетаемости той или иной омонимичной лексической единицы. Данный вопрос требует глубокой проработки и поэтапного решения и является, пожалуй, темой нескольких кандидатских диссертаций.

Что касается морфологических омонимов, то можно с уверенностью утверждать, что использование контекстного подхода обязательно даст результаты. Приведем несколько наблюдений за работой автоматического морфологического анализатора.

1. Если после токена со множественными разборами идет послелог (только в случае, когда токен получает однозначную характеристику как POST), то можно частично или в некоторых случаях полностью снять омонимию: послелог может сочетаться только с именными частями речи. Правило: PTCPL|N|NUM|ADJ|PRON + POST, при этом нужно оговориться: в случае если причастие или прилагательное субстантивируется, т. е. перестает быть атрибутивной формой глагола или именем прилагательным, а становится существительным, выполняя его функции и принимая на себя все характеристики, прежде всего становится изменяемым. Невозможно сочетание CONV|V|PART|CONJ|INJ + POST.

2. Если линейная последовательность состоит из двух токенов, первый из которых получает после анализа программы множественный разбор как N.NOM/ACC и ADJ, а второй — однозначный разбор как N, но множественный как NOM/ACC, то можно полностью снять омонимию, поскольку невозможно сочетание двух существительных в именительном или винительном немаркированном падеже, за исключением однородных членов предложения, соединенных как бессоюзной[1], так союзной связью (*болн* ‘и’, *хойр* ‘и’).

3. В сочетании типа *эмгн өвгн хойр* ‘старик и старуха’ токен *хойр* может разбираться как числительное ‘два’. В конструкциях N + N + *хойр* последняя лексическая единица может получать единичный разбор как CONJ, поскольку занимает постпозицию по отношению к слову, к которому оно относится. В конструкции *хойр* + N, напротив, первый элемент находится слева, т. е. в препозиции. Следовательно, учитывая все положение соседнего окружения, можно провести разграничение.

[5] В данном случае необходимо учитывать знаки пунктуации.

4. Лексическую омонимию можно снять при помощи фильтров по семантическим критериям. Если заранее известны списки омонимов, то можно предположить, что каждый из омонимов имеет свои правила сочетаемости с другими словами. Например, возьмем существительные-омонимы. Формальным показателем является контекст, т. е. правое или левое окружение. В случае с левым окружением омоним может иметь определение, которое выражено прилагательным или порядковым числительным. Существуют ограничения на сочетаемость разных тематических групп определений.

Омоним	**Перевод**	**Тематическая группа омонима**	**Тематическая группа определения**
ааг[1]	‘высокомерие; гордость’	качество человека	определения, обозначающие 1) признак предмета через его отношение к социальной группе, 2) поведение, 3) оценку
ааг[2]	‘настой (чая; лекарства)’	жидкость	определения, обозначающие 1) признак предмета через его отношение к растениям, 2) физические свойства (насыщенность, крепость, цвет и др.)
ааг[3]	‘язычок (у рыболовного крючка); зазубрина остроги’	часть инструмента	определения, обозначающие 1) наличие или отсутствие содержания, 2) физические свойства (форма) 3) признак предмета через его отношение к материалу, из которого он сделан
ааг[4]	‘развилка’	пространство	определения, обозначающие 1) очередность, 2) модальность (подлинная, ложная и т. д.)

Теоретически если задать подобного рода фильтры на сочетаемость, то можно частично снять омонимию. Можно использовать и сочетаемость данных слов с глаголами и их формами в целях повышения точности снятия омонимии.

Наблюдения за результатами разбора и выведение правил калмыцкого языка в целях решения проблемы лексико-грамматической неоднозначности во многом будут способствовать разработке снятия омонимии для ряда агглютинативных языков, к которым относятся родственные калмыцкому бурятский и монгольский языки. На наш взгляд, разработка гибридного подхода может дать более высокие результаты, чем при работе алгоритма, основанного только на одном методе (или детерминированном, или вероятностном).

В первую очередь, наряду с анализом окружения токена с множественным разбором, статистических данных об этом окружении, необходимо провести снятие омонимии вручную для создания задела корпуса текстов для проверки выработанных правил в снятии неоднозначности. В этот корпус должны быть включены тексты разных стилей, разного времени создания (начиная со второй половины XX в.[1]) и разных авторов. Объем текстотеки должен быть небольшой — до 1 млн словоупотреблений. Когда массив данных корпуса со снятой омонимией будет создан, только после этого можно приступать к разработке модуля. На данный момент отсутствует возможность проверить свои эмпирические наблюдения на небольшом объеме текстов, поскольку в корпусе нет раздела со снятой омонимией. Главной задачей на этом, первом, этапе было создание репрезентативной текстотеки на калмыцком языке и первичного задела по разработке программного обеспечения работы корпуса.

В частотном словаре современного калмыцкого языка, который сейчас разрабатывается в Калмыцком институте гуманитарных исследований РАН, не разграничивается лексико-грамматическая омонимия, поскольку создание модуля по снятию омонимии дело будущего. Другими словами, одной единицей считаются слова *он*[1] ‘год; годовщина’, *он*[2] ‘насечка, зазубрина’, *көвәд* N.Dat=‘берег’ и *көвәд*[1]

[6] Время создания мы сознательно ограничиваем по двум причинам: во-первых, до указанного времени письменность была основана на латинице и кириллице без четко установившейся орфографической нормы, во-вторых, с этого периода была выработана орфографические правила в калмыцком языке, хотя тексты свидетельствуют, что нормы не были прочно закреплены на практике (большое количество разновариантных написаний слов, не принадлежащих к диалектам).

Conv.Ant=‘заставить плавать на воде’/ *көвәд*[2] Conv.Ant=‘заставить линять’. Однако было принято решение о необходимости публикации «верхушек» частотных списков лемм, словоформ, граммем[1] по причине того, что данные можно использовать при построении методики преподавания калмыцкого языка как иностранного с подачей лексического материала с учетом частоты слов. К тому же языковая ситуация, сложившаяся в Республике Калмыкия, когда родной язык как средство коммуникации, не говоря уже как о средстве мышления, когниции, находится в процессе утраты, вынуждает предпринимать меры по сохранению языка. Как известно, без теоретической основы невозможно выстроить продуктивную методику преподавания языка.

Одним из приложений словаря будет таблица омонимичных форм с указанием частоты в текстах, хотя мы оговариваем, что картина статистики употребления каждой отдельной лексемы в ранговом соотношении будет немного искаженной на 1–2 пункта, однако на абсолютную частоту это не повлияет. По таблице омонимов, тем не менее, можно сравнить их частоту с другими лексемами и сделать вывод о частоте их употребления в речи.

Выводы

Таким образом, снятие омонимии безусловно является фундаментальной проблемой и требует глубокой и детальной проработки, от решения которой зависит частота правильного выполнения пользовательских запросов, уменьшения хранения информации о том или ином токене и, следовательно, быстродействия системы и мн. др.

Литература

Биткеев П. Ц. Проблемы фонетики калмыцкого языка (Квантитативные и квалитативные изменения гласных). Элиста: Калм. кн. изд-во, 1975. 170 с.

Зеленков Ю.Г., Сегалович И. В., Титов В.А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов [электронный ресурс] // URL: http://www.dialog-21.ru/Archive/2005/Zelenkov%20Segalovich/Zelenkov_Segalovich.htm (дата обращения: 20.09.2014).

[7] Полные частотные списки планируется опубликовать в электронном издании словаря.

Калмыцко-русский словарь / под ред. Б. Д. Муниева. М.: Русский язык, 1977. 768 с.

Куканова В.В., Бембеев Е.В., Мулаева Н.М., Очирова Н.Ч. Национальный корпус калмыцкого языка: архитектура и возможности использования // Вестник Калмыцкого института гуманитарных исследований РАН. 2012. № 3. С. 138–150.

Куканова В.В., Каджиев А.Ю. Алгоритм работы морфологического парсера калмыцкого языка // Писменото наследство и информационните технологии [Текст]: Материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В.А. Баранов, В. Желязкова, А.М. Лаврентьев. София; Ижевск, 2014. С. 116–119.

Порохнин А.А. Анализ статистических методов снятия омонимии в текстах на русском языке // Вестник Астраханского технического университета. Серия: Управление, вычислительная техника и информатика. 2013. № 2. С. 168–174.

Трофимова С. М., Рассадин В. И. О недостатках действующей калмыцкой орфографии // Актуальные проблемы диалектологии языков народов России: матлы XIII международной конференции. Уфа, 2013. С. 131–132.