В. В. Куканова

К постановке проблемы создания обратного словаря калмыцкого языка (предварительные замечания)

В настоящее время пристальное внимание лингвистов разных стран направлено на создание корпусов языков, под которым и понимается «информационно-справочная система, основанная на собрании текстов на некотором языке в электронной форме» [1], и на решение проблем, связанных с автоматической обработкой многомиллионного текстового массива. Информационно-справочная база данных по некоторому языку предполагает разметку материала, причем это аннотирование может относиться как ко всему тексту (т. н. метаразметка), так и его элементам (на уровне фонемы, морфемы, слова, предложения и текста как структуры). Соответственно выделяют морфемную, морфологическую, синтаксическую, семантическую и другие разметки, вследствие подобного аннотирования облегчается поиск необходимого элемента в массиве текстов. Если на сбор материала для исследования некоторой проблемы уходили месяцы (а то и годы), то в этом случае требуется всего несколько секунд. Работа по разметке, проделанная однажды одним лингвистом, находится в открытом доступе (on-line), т. е. другой исследователь может воспользоваться результатами этой работы, соблюдая все этические нормы. Корпус позволяет изучать, например, лексический строй языка с опорой на тексты, тем самым реализуется современный подход к исследованию языка – корпусный, который учитывает не просто отдельные примеры, а все контексты с данной лексемой, в результате чего перед исследователем полное описание употребления того или иного слова, его значений, свойств, дистрибуции и т. д. Другими словами, корпус является прекрасным инструментом для организации и моделирования научных исследований. Создание корпуса откроет новые горизонты в калмыцкой лингвистической науке, как например, создание частотного, толкового словаря калмыцкого языка, словарей сочетаемости и многое другое.

Быстрое развитие корпусной лингвистики связано, конечно, с распространением и постоянной модернизацией компьютерных технологий, и в данный момент создано огромное количество корпусов и баз данных тех или иных языков [2; 3; 4; 5; 6; 7; 8 и т. д.]. В большинстве случаев объектом для создания корпусов выступали языки, которые являются международными, как русский, английский, немецкий и др. Очевидна актуальность создания подобного корпуса и для исчезающих языков, в частности для калмыцкого, поскольку это еще один способ сохранения культурного наследия и этнического своеобразия народа, выражающегося в языке, который сам находится на грани исчезновения. Напомним, что язык является главным признаком этнической идентичности нации.

Говорить о малоисследованности калмыцкого языка как такового в структурном или функциональном плане, так и в сопоставлении с типологически родственными (монгольским, бурятским и др.) и разноструктурными языками не приходится. В настоящее время сотрудниками Калмыцкого института гуманитарных исследований РАН собираются тексты на калмыцком языке, принадлежащие разным стилям и жанрам, а также различным периодам функционирования языка.

Для реализации данного фундаментального проекта необходимо решить ряд важных задач как теоретического, так и практического характера. Одной из таких задач является создание обратного словаря калмыцкого языка. Поясним, что значит обратный, или инверсионный, словарь, потому что, как пишет А. А. Зализняк, подобные словари необычны в лингвистике [9, с. 3]. Если в толковом, орфографическом и др. словарях расположение лексических единиц алфавитное слева направо, то в инверсионных — тоже алфавитное, но только справа налево. Например:

‘рожковый’	*аагта*
‘еще пригодный’	*тангта*
‘сомнительный, неопределенный’	*маһдта*
‘пряник’	*балта*
‘смолистый’	*салмта*
‘поручительский; ответственный’	*дааврта*
‘замужняя’	*залута*
‘тридцатилетний’	*һучта*

Заметим, что в обратном словаре рядом могут находится не все слова одного грамматического разряда, как, например, среди прилагательных в списке, приведенном выше, можно обнаружить одно существительное *балта* (также как и в русском языке). По мнению и опыту А. А. Зализняка, таких случаев встречается в русском языке немного [9, с. 4].

Подобный словарь необходим для выявления словоизменительных разрядов, классов разных частей речи. Как считает А. А. Зализняк, слова, оканчивающиеся одинаково, должны иметь одинаковые парадигмы словоизменения [9, с. 4]. Данные словоизменительные разряды и классы в свою очередь требуются для создания лемматизатора – специальной программы, приводящей словоформу к начальной форме, или лемме, которая позволит автоматически обрабатывать массивы текстов, хотя, естественно, снимать омонимию придется вручную или при помощи отдельных модулей-приложений. Однако это следующий этап в работе по созданию корпуса калмыцкого языка.

Сам обратный словарь можно использовать и для выборки материала того или иного исследования, например, если лингвиста интересуют прилагательные со словообразовательным суффиксом *-ma/mə*, то можно по обратному алфавиту с легкостью найти данные единицы и проанализировать различные морфонологические процессы, происходящие на стыке корня и суффикса. О других преимуществах и возможностях применения обратного словаря см. работу Р. В. Бахтуриной и И. А. Мельчука [10].

Зарождение идеи инверсионных словарей восходит к XIII–XIV в., т. е. к эпохе средневековья, хотя создавались они для практических целей – для написания стихотворений. В лингвистических целях подобные словари стали использоваться на рубеже XIX и XX в., когда активизировался интерес к исследованию «мертвых» языков (санскрита, латинского и др.). Что касается русского языка, то существует несколько инверсионных словарей, это, прежде всего, словарь под редакцией Г. Бильфельдта [11] и, конечно, знаменитый обратный словарь А. А. Зализняка [12]. Последний явился поворотным моментом в развитии отечественных обратных словарей и, шире, словарей любого толка, так как обработка лексических единиц производилась с помощью вычислительной техники, что во многом упростило работу лингвистов.

В лингвистике создание первых обратных словарей для урало-алтайских языков приходится на конец 1960-х и 1970-е гг., наиболее разработанной частью является тюркское направление, о чем свидетельствует значительное количество словарей для тюркских языков (узбекского [13], казахского [14], татарского [15] и башкирского [16]). Надо сказать, что методика обработки единиц, в частности, для монгольских языков еще не выработана полностью, за исключением обратного словаря монгольского языка [17] и работ С. А. Крылова (см., например) [18]. К сожалению, попытки создания подобного словаря не предпринимались ни в калмыцкой, ни в бурятском языкознаниях.

В качестве образца для создания инверсионного словаря калмыцкого языка используется словарь русского языка под редакцией А. А. Зализняка

и обратный словарь монгольского языка. Однако необходимо помнить, что русский язык является флективным, а калмыцкий язык – классическим агглютинативным языком по своей структуре, где каждый структурный элемент слова несет свое собственное грамматическое значение, т. е. граммемы падежа, числа и др. выражаются в отдельных формантах. Например: в слове *школ-муд-ын* первый формант *-муд* выражает граммему множественного числа, а второй *-ын* – граммему родительного падежа. Гипотетически достаточно легко будут определяться граммемы словоформ с помощью лемматизатора, хотя трудности будут возникать при различных изменениях на стыке формантов внутри слова.

По мнению некоторых лингвистов, агглютинативные языки рассматриваются как языки с традиционно бедной морфонологией [19]. Исследования С. А. Крылова на примере халха-монгольского языка доказали, что агглютинативные языки обладают богатыми морфонологическими процессами, начиная с различных видов фузии и заканчивая явлениями супплетивизма [18]. Поскольку халха-монгольский и калмыцкий языки родственные, то можно предположить, что различные морфонологические процессы на стыке морфем характерны и для калмыцкого языка. К тому же «…морфонология, являющаяся <…> связующим звеном между фонетикой и морфологией, призвана благодаря такому своему положению в системе грамматического описания дать всеобъемлющую характеристику каждого языка. Возможно, что при установлении языковых типов с морфонологических позиций как раз и откроется возможность для создания рациональной типологической классификации языков земного шара» [20, с. 119].

Не стоит забывать, что русский и калмыцкий языки имеют несколько иную морфологию, хотя частеречная структура языков почти одинакова, за исключением послелогов и предлогов, по своей сути, являющихся одним и тем же. Как справедливо отмечает А. А. Зализняк, для составления грамматического словаря русского языка имеет только значение, изменяема или не изменяема та или иная единица, и этого было бы достаточно для определения парадигм, так как почти во всех частях речи русского языка можно обнаружить слова, которые не изменяются. Например, среди существительных класс так называемых несклоняемых: *шоссе*, *кино*, *пальто* и др.; среди прилагательных – несколько цветовых слов: *хаки*, *беж*, *бордо*. Однако для калмыцкого языка, видимо, важно знать, какой части речи принадлежит та или иная лексема, поскольку, например, все слова-существительные изменяются по числам и падежам. Оговоримся, что существительные Singularia Tantum и Pluralia Tantum, вслед за А. А. Зализняком, несут потенциальную возможность образования своей грамматической пары и могут иметь полную парадигму. Думается, что

необходимо выделять и грамматические разряды, и части речи применительно к калмыцкому языку, так как в каждом своем случае необходима и та, и другая информация. Прежде всего можно выделить изменяемые и неизменяемые грамматические разряды в калмыцком языке (см. таблицу), так как для правильной работы программы-лемматизатора необходимо соотношение: один формальный компонент и одно значение, только в этом случае отсутствуют проблемы с омонимичными лексическими и грамматическими формами.

Что касается имени прилагательного, то данная часть речи неизменяема в калмыцком языке: не образует словоформ и степеней сравнения. Но если прилагательное субстантивировалось, то оно получает все грамматические признаки, присущие существительному. Явление субстантивации прилагательных широко распространено в калмыцком языке, и правил перехода из одной части речи в другую не существует (или существуют, но пока еще не исследовались в калмыцком языкознании), поэтому каждое калмыцкое прилагательное – это потенциально изменяемое слово. Например, *би бориг үзҗәнәв* ‘я вижу гнедого’.

Сопоставление частей речи и словоизменения в калмыцком и русском языках

№	Часть речи	Калмыцкий язык	Русский язык
1.	Имя существительное	+	+*
2.	Имя прилагательное	– (?)	+*
3.	Наречие	–	–
4.	Числительное	+	–*
5.	Местоимение	+	+*
6.	Глагол	+	+
7.	Послелог/Предлог	–	–
8.	Союз	–	–
9.	Частицы	–	–
10.	Междометие	–	–

Астериксом обозначены те части речи, в которых имеются неизменяемые слова, т. е. парадигма словоизменения этих единиц представлена омонимичными формами. Существуют также неполные парадигмы, в которых отсутствуют некоторые элементы: так, например, местоимение *эврән* имеет только граммему именительного падежа.

Другой проблемой теоретического плана является так называемые двойное и возвратное склонения в калмыцком языке. Имена существительные, оформленные одним из падежных аффиксов простого склонения,

могут принимать дополнительные аффиксы. Двойные падежи образуются путем присоединения двух падежных окончаний к основе одного и того же имени, где первый выполняет словообразовательную функцию, а второй – словоизменительную. Имя существительное калмыцкого языка в двойном склонении изменяется не по всем падежам (основой для него могут выступать словоформы в родительном, совместном и дательном падежах). Возвратное склонение характеризуется тем, что к имени присоединяется сначала словоизменительный аффикс, а потом уже возвратная частица. Парадигма двойного и возвратного склонений неполная, так как отсутствуют формы множественного числа, кроме того, существуют семантические ограничения в изменении существительного по двойному и возвратному склонениям, которые предстоит еще выяснить.

Рассмотрим некоторые проблемы практического характера, связанные с созданием обратного словаря. Данные вопросы требуют незамедлительного решения не только в целях создания инверсионного словаря, но и для калмыцкого языкознания в целом.

Во-первых, проблема обозначения так называемых редуцированных гласных в калмыцком языке, имеющая важнейшее значение для теории и практики преподавания калмыцкого языка в школах. Дети в процессе обучения калмыцкому языку, по сути, как иностранному не могут интуитивно определить правила, где необходимо произносить редуцированные, что в свою очередь создает некоторые трудности в произношении сочетаний согласных в калмыцких словах. Что касается теоретической грамматики, то при обозначении редуцированных гласных будет легко распознать позиции, где происходит чередование гласных с отсутствием звука вообще при словообразовании и формообразовании. Так, например, слово *тоосн* [тӧсън] ‘пыль’ – *тоосна* [тӧсна] ‘пыли’ (Gen.); ‘пыльный’. Вопрос о редуцированных гласных требует специальных исследований, отвечающих современным достижениям в области фонетики. В результате подобного анализа будет определен фонемный статус «неясных» гласных звуков, позиций, где они появляются и где отсутствуют. Если будет доказано, что редуцированные – это самостоятельные фонемы, выполняющие смыслоразлитель ные функции в слове, то станет вопрос об их обозначении. Для автоматической обработки чрезвычайно важно иметь «на входе» формально выраженные данные.

Во-вторых, предстоит решить вопрос о структуре словарной статьи, а также списке слов, которые необходимо обрабатывать, о программе, где будут обрабатываться лексические единицы. Список слов, изъятых из Калмыцко-русского словаря [21], на данный момент составляет около 26 тыс. единиц, что является лишь половиной от необходимого минимума,

поэтому одной из важнейших задач считается пополнение списка лексических единиц. Для осуществления этой задачи требуется сканировать и распознавать тексты на калмыцком языке, а потом уже извлекать вручную языковой материал в виде списка словоформ, поскольку лемматизатора для приведения словоформ в начальную форму в калмыцкой прикладной лингвистике пока не существует. К тому же словарь должен быть кумулятивным, т. е. обладать способностью постоянно пополняться.

Думается, что создание обратного словаря калмыцкого языка является приоритетной областью в теоретическом и прикладном языкознании и необходимо для фундаментальных исследований в русле той или иной лингвистической школы. Опыт и методику, выработанную в ходе работы над словарем, могут применять к другим языкам, имеющим агглютинативную структуру.

ЛИТЕРАТУРА

1. Что такое корпус? [Электронный ресурс] // URL: http://www.ruscorpora.ru/corpora-intro.html (15.04.2011).
2. ХАНКО – Хельсинкский аннотированный корпус [Электронный ресурс] // URL: http://www.ling.helsinki.fi/projects/hanco/ (15.04.2011).
3. Национальный корпус русского языка [Электронный ресурс] // URL: http://ruscorpora.ru/ (15.04.2011).
4. Банк английского языка [Электронный ресурс] // URL: http://www.collins.co.uk/Corpus/CorpusSearch.aspx (15.04.2011).
5. Британский национальный корпус [Электронный ресурс] // URL: http://sara.natcorp.ox.ac.uk/ (15.04.2011).
6. Корпус современного китайского языка (LIVAC Synchronous Corpus) [Электронный ресурс] // URL: http://www.cilta.unibo.it/ricerca.htm (15.04.2011).
7. Корпус современного итальянского языка CORIS/CODIS [Электронный ресурс] // URL: http://www.cilta.unibo.it/ricerca.htm (15.04.2011).
8. Мангеймский корпус немецкого языка (Institut für Deutsche Sprache, Mannheim, Germany) [Электронный ресурс] // URL: http://corpora.ids-mannheim.de/~cosmas/ (15.04.2011).
9. Зализняк А . А. Предисловие // Зализняк А. А. Грамматический словарь русского языка: Словоизменение: Около 100 000 слов. 3-е изд., стереотип. М.: Русский язык, 1987. С. 3–10.
10. Бахтурина Р. В., Мельчук И. А. Рец.: M. L. Alinei. Dizionario inversa italiano. Con indici e liste di frequenza delle terminazioni. The Hague, Mouton and Go, 1962. 607 с. // Вопросы языкознания. 1965. № 5. С. 128–133.
11. Bielfeldt H. H. Rucklaufiges Worterbuch der russischen Sprasche der Gegenwart. Berlin, 1958. 392 p.

12. Зализняк А. А. Грамматический словарь русского языка: Словоизменение: Около 100 000 слов. 3-е изд., стереотип. М.: Рус. яз., 1987. 880 с.
13. Кунгуров Р. К., Тихонов А. Н. Обратный словарь узбекского языка. Самарканд, 1968. 187 с.
14. Бектаев К. Б. Обратный словарь казахского языка. Алма-Ата, 1971. 288 с.
15. Ахтямов М. Х. Обратный словарь татарского языка = Татар телнең кире сүзлеге: Ок. 32000 слов. Уфа: Башкир. ун-т, 1999. 196 с.
16. Ахтямов М. Х. Обратный словарь башкирского языка = Башҡорт теленең кире һүзлеге: Ок. 21000 слов. Уфа, 1999. 236 с.
17. Болд Л. Орчин цагийн монгол хэлний тонгоруу толь. (Обратный словарь современного монгольского языка). Улаанбаатар, 1976. 236 с.
18. Крылов С. А. Теоретическая грамматика современного монгольского языка и смежные проблемы общей лингвистики. Ч. 1. Морфемика. Морфонология. Элементы фонологической трансформаторики (в аспекте общей теории морфологических и морфонологических моделей). М.: Вост. лит., 2004. 479 с.
19. Грунтов И. А. Рец. на кн.: Крылов С. А. Теоретическая грамматика современного монгольского языка и смежные проблемы общей лингвистики. Ч. 1. Морфемика. Морфонология. Элементы фонологической трансформаторики (в аспекте общей теории морфологических и морфонологических моделей) // Вопросы языкознания. 2006. № 1. С. 148–150.
20. Трубецкой Н. С. Некоторые соображения относительно морфонологии // Пражский лингвистический кружок. М., 1967. С. 115–119.
21. Калмыцко-русский словарь / под ред. Б. Д. Муниева. М.: Русский язык, 1977. 768 с.