УДК 81’373.47
ББК 81.23

ЧАСТОТНЫЙ СЛОВАРЬ СОВРЕМЕННОГО КАЛМЫЦКОГО ЯЗЫКА: ПРАВИЛА АНАЛИЗА ТЕКСТОВОГО МАТЕРИАЛА

Frequency Dictionary of Modern Kalmyk Language: Rules of Analysis of Text Material

Е. В. Бембеев (E. Bembeev)[1], *В. В. Куканова (V. Kukanova)*[2], *А. Ю. Каджиев (A. Kadzhiev)*[3]

[1] кандидат филологических наук, старший научный сотрудник Лаборатории прикладной и экспериментальной лингвистики Калмыцкого института гуманитарных исследований Российской академии наук (Ph.D. of Philology, Senior Reseacher of the Applied and Experimental Linguistics Laboratory at the Kalmyk Institute for Humanities of the Russian Academy of Sciences). E-mail: galdma@yandex.ru.

[2] кандидат филологических наук, заведующий Лабораторией прикладной и экспериментальной лингвистики Калмыцкого института гуманитарных исследований Российской академии наук (Ph.D. of Philology, Head of the Applied and Experimental Linguistics Laboratory at the Kalmyk Institute for Humanities of the Russian Academy of Sciences). E-mail: vika.kukanova@gmail.com.

[3] инженер-исследователь Лаборатории прикладной и экспериментальной лингвистики Калмыцкого института гуманитарных исследований Российской академии наук (Research Engineer of the Applied and Experimental Linguistics Laboratory at the Kalmyk Institute for Humanities of the Russian Academy of Sciences). E-mail: arasha.kadzhiev.work@gmail.com.

Статья посвящена описанию правил анализа текстового материала для создания частотного словаря калмыцкого языка на материале Национального корпуса калмыцкого языка (www.kalmcorpora.ru), который состоит из художественных текстов второй половины XX – начала XXI в., а также газетных статей и расшифровок устной речи. Объем художественных (прозаических и поэтических) текстов превышает 10 млн словоупотреблений. Тексты в корпусе, а также отдельные элементы текста (словоформы, знаки препинания, абзацы и т. п.) особым образом аннотированы. Создаваемый частотный словарь калмыцкого языка будет носить пилотный характер, поскольку это первый опыт разработки словаря подобного типа. На наш взгляд, объем созданного корпуса калмыцкого языка позволяет описать язык с точки зрения частотности употребления языковых единиц и значений: словоформ, слов, конструкций (2- и 3-граммных), грамматических значений, букв и др.

Ключевые слова: корпусная лингвистика, квантитативные методы в лингвистике, частотный словарь, калмыцкий язык, правила лемматизации.

The article is devoted to description of the rules for text material analysis for creating the Frequency Dictionary of the Kalmyk language on the basis of the National Corpus of the Kalmyk Language (www.kalmcorpora.ru) which includes the texts of the literary works published in the second half of the 20th and at the beginning of the 21st centuries as well as newspaper articles and transcripts of spoken language. The volume of the fiction (prose and poetry) exceeds 10 mln. words. The texts in the Corpus as well as certain elements of the texts (word-forms, punctuations signs, paragraphs, etc.) have special annotations. The Frequency Dictionary created on the basis of the Corpus is a pilot model as it is the first attempt to develop a dictionary of this type. In our opinion, the size of the created Corpus of the Kalmyk Language allows to describe the language from the point of view of usage frequency of language units and meanings: word-forms, words, constructions (2 and 3-gramms), grammatical meanings, letters, etc.

In 2013, the experimental version of the National Corpus of the Kalmyk Language was launched, but it did not have any morphological and semantic annotations though the closed data had already possessed these types of annotations. The material containing the annotations will be open after the analyzer’s program code will be adjusted, and its efficiency will reach 90%. At the present moment, the model of the algorithm of work of the morphological parser for the Kalmyk language successfully analyzes 70% of any text providing only unambiguous parsing at the same time. About 20% of the texts have multitude possible variants of automated analyses, though 10% of the texts have no parsing as there are no stems for them in the dictionary (they are mostly Russian loanwords which were not included into the Dictionary edited by B.D. Muniev [1977] and some proper names).

The main idea of developing the Frequency Dictionary is that the most frequently used language units are the most significant ones in any language but at the same time non-frequent elements are of the same significance but from the other point of view. They can carry some traces of historical development and can belong to various terminological systems which evidences that a lexical unit is out of use in speech.

The issue of the language units and meanings frequency is not developed in the Kalmyk linguistics that is why for researching the frequency characteristics of the Kalmyk speech one should first of all identify and justify the parameters for distinguishing frequency and describing frequency characteristics of the Kalmyk speech. Thus the aim of this article is to describe the rules for analyzing lexical units in order to develop the Frequency Dictionary of the Kalmyk language where the observation unit is a lemma - that is an initial form of the language without its lexical and grammatical annotations. However, it does not mean that the dictionary development will not take into account the Kalmyk grammar: processing of word-forms and working out lemma vocabulary are regulated by the rules of the formalized description of the Kalmyk language grammar, besides for each part of speech there is a separate description. The main and basic issue is to define the boundaries for the notions of a word and a lemma (an initial form of a word).

The article provides the rules for textual material analysis in order to create the Frequency Dictionary of the Kalmyk language. These rules are built on the principles for developing "The Frequency Dictionary of the Russian Language" [Frequency Dictionary … 1977] and "The Grammar Dictionary of the Russian Language" [Zalizniak 1987] which were revised for the purposes of the Kalmyk language, while for the units which do not exist in the literary written language the rules have been developed anew. Each part of speech has its own set of rules which regulates the work of the morphological parser to process lineal letter sequence of the vocabulary element for the Frequency Dictionary.

Keywords: Corpus Linguistics, quantitative methods in Linguistics, frequency dictionary, the Kalmyk language, the rules for lemmatization.

В последние годы с развитием информационных технологий становится легче и быстрее создавать частотные словари, в которых эксплицирована структурно-вероятностная модель того или иного языка. Методы квантитативной лингвистики приобретают все больший интерес среди исследователей, поскольку результаты количественной обработки текстов можно применить в решении не только прикладных задач, но и фундаментальных теоретических проблем. Частотный словарь «…включает в себя упорядоченный список слов или других языковых единиц (словоформы, словосочетания), которые зарегистрированы составителем в обследованном им тексте, фрагменте текста или корпусе текстов и снабжены данными о частоте их употребления в тексте (речи). С его помощью можно попытаться ответить на вопросы: как много слов в языке (тексте), с какой интенсивностью они используются в речи, какие из них предпочтительнее в той или иной сфере коммуникации у того или иного автора и т. д.» [Долинский 2004: 285].

Создание частотных словарей на материале русского языка имеет уже продолжительную историю, начиная с 1950-х гг. [см., например, Лённгрен 1993; Степанова 1976; Частотный словарь … 1977]. Венцом развития отечественной квантитативной лингвистики стал, конечно, Частотный словарь, основанный на материале Национального корпуса русского языка [Ляшевская, Шаров 2009], который насчитывал на момент работы над словарем 100 млн словоупотреблений[1].

Отметим, что в калмыцком языкознании еще ни разу не предпринимались попытки компилирования частотных словарей, поскольку, во-первых, отсутствовал репрезентативный объем оцифрованных текстов на калмыцком языке; во-вторых, развитие компьютерных технологий и уровень их применения не позволяли этого сделать. Появление частотного словаря калмыцкого языка сыграло бы определенную роль в аспекте сохранения языка.

В 2013 г. была запущена тестовая версия Национального корпуса калмыцкого языка без морфологической и семантической разметки (http://www.kalmcorpora.ru), хотя данный тип аннотации был осуществлен в закрытой базе данных. Программный код морфологического анализатора еще не совершенен, требует «отладки» и доведения его работы до 90 %. В настоящем виде модель алгоритма работы морфологического анализатора калмыцкого языка успешно анализирует 70 % текста и выдает при этом однозначный разбор. Около 20 % текста имеют множественные вероятностные варианты автоматического анализа. У 10 % вообще отсутствуют разборы ввиду того, что в словаре основ нет их стемов (в основ-

[1] Ср. с частотным словарем под ред. Л. Н. Засориной, который основан на текстах общим объемом 1 млн словоупотреблений [Частотный словарь … 1977].

ном, это слова из русского языка, не вошедшие в Калмыцко-русский словарь под ред. Б. Д. Муниева [1977], русские собственные имена, орфографические варианты и др.).

Национальный корпус калмыцкого языка состоит из художественных текстов второй половины XX – начала XXI в., а также газетных статей и расшифровок устной речи. Объем художественных (прозаических и поэтических) текстов превышает 10 млн словоупотреблений. Тексты в корпусе, а также отдельные элементы текста (словоформы, знаки препинания, абзацы и т. п.) особым образом аннотированы [см. подробнее: Куканова и др. 2012а; 2012б]. Разрабатываемый частотный словарь калмыцкого языка будет носить пилотный характер, поскольку это первый опыт разработки словаря подобного типа. На наш взгляд, объем созданного корпуса калмыцкого языка позволяет описать язык с точки зрения частотности употребления языковых единиц и значений: словоформ, слов, конструкций (2- и 3-граммных), грамматических значений, букв и др.

Актуальность создания частотного словаря несомненна. Во-первых, частотный словарь позволит определить границы лексической системы, которая имеет свое ядро и периферийные поля (т. е. частотные и нечастотные элементы). Создание частотных списков для калмыцкого языка необходимо и в плане исследований общей типологии языков. В аспекте практической значимости создания частотного словаря можно говорить о решении прикладных задач распознавания, усовершенствования орфографии и др. С наиболее частотных единиц, как правило, начинается обучение языку, объясняется, каково их значение и как использовать их в речи. К тому же наиболее частотные слова обладают разветленной системой значений, нерегулярной морфологией, широким идиоматическим функционированием. Большинство словарей, предназначенных для изучения того или иного языка, имеет в словарной статье помету о частотности.

Главная идея создания частотного словаря заключается в том, что наиболее частотная единица является наиболее важной в системе и в то же время нечастотные элементы занимают уникальное место в лексической системе. Они могут содержать следы исторического развития, принадлежать той или иной терминологической системе, что свидетельствует о неупотребительности лексической единицы в речи.

Как известно, частотные словари составляются с опорой на различные единицы счета: словоформы, лексемы (с различением или неразличением разных типов омонимов), словосочетания, грамматические значения. Обычно за единицу словника принимается либо словоформа, либо лексема. В качестве единицы словника может выступать и граммема [Крылов 2013]. Выбирая в качестве единицы счета словоформу, составитель словаря опирается только на графическую эквивалентность, никакого морфологического анализа текста не производится. Если в качестве единицы количественной обработки брать лемму, то в создании репрезентативного частотного словаря не обойтись без автоматического анализа текста.

Вопрос о частотности языковых единиц и значений является не разработанным в калмыцком языкознании, поэтому для исследования частотных характеристик калмыцкой речи следует первоначально определить и обосновать параметры анализа частотных характеристик калмыцкой речи. Целью данной статьи и является описание правил анализа лексических единиц в свете создания частотного словаря современного калмыцкого языка, где единицей счета выступает лемма, т. е. исходная форма слова, без сопровождения лексико-грамматических помет. Однако это не означает, что словарь будет строиться без учета грамматики калмыцкого языка: обработка словоформ и создание словника лемм регулируется правилами формализованного описания грамматики калмыцкого языка, причем правила выводятся для каждой части речи отдельно. Главными и основополагающими вопросами являются определение границы слова и понятие леммы (начальной формы слова).

Проблема границ слова — один из нерешенных вопросов лингвистики, на который до сих пор нет точного и однозначного ответа. Делимитация слова в речевой цепи зависит от целей исследования и наличия у исследователей возможностей программно обработать линейную последовательность. Например, в прикладной лингвистике используется графический подход: слово определяется как последовательность знаков, ограниченная пробелами [Гак 1990; Касевич 1977: 57–58]. Этот подход удобен для автоматической обработки текстов, так как графический анализатор сегментирует слова по пробелам, программа понимает

данный знак как делиметр. Однако в языках существуют «сочетания» нескольких графических слов, которые, по сути, являются одним словом, несмотря на то, что его компоненты пишутся отдельно или через дефис. Каждый из этих элементов обладает собственным ударением или сочетанием главного и побочного ударений. В целом оно представляет собой одно лексическое значение. Это так называемые сложные слова, или композиты (компаунды — compound words[1]), которые изобилуют в тюркских и монгольских, а также в английском, испанском, немецком языках.

Что касается калмыцкого языка, то в нем также существует множество компаундов. Для получения чистой статистики, конечно, следует учитывать в качестве единицы счета не токен, а слово, которое может состоять из двух и более токенов. На данном этапе программа TextAnalyzer, созданная в Калмыцком институте гуманитарных исследований РАН, вычленяет сложные слова, написанные через дефис, а также редупликации. С компаундами, которые пишутся через пробел, дело обстоит немного сложнее. Сейчас модуль выделения сложных слов разрабатывается, в идеале эта часть программы должна обнаруживать и те слова, которые пишутся через пробел. По этой причине для нас это еще не решенный вопрос. На данный момент мы пока примем в качестве счета токен, т. е. слово, разграниченное пробелами, без учета компаундов.

Другой проблемой является определение леммы (начальной формы). Что касается именных частей речи, здесь нет теоретически не разрешимых проблем. Как известно, в калмыцком языке отсутствует инфинитив, хотя некоторые исследователи считают, что формы на *-х* не выражают значения времени, наклонения, лица и числа, т. е. являются инфинитивом. Т. А. Бертагаев рассматривает инфинитив как особую глагольную форму, которая не является атрибутивом имени и не склоняется [Бертагаев 1964: 41]. В этом вопросе мы разделяем академическую точку зрения: в монгольских языках отсутствует инфинитив. Традиционно в словарях в качестве заголовочного слова дается форма причастия в будущем времени, поэтому именно причастие в будущем времени приводится в качестве леммы слова.

Таким образом, единицей описания при создании словника частотного словаря современного калмыцкого языка признается не только графическое слово (текстоформа, т. е. «от пробела до пробела»), с которым «работают» многие морфологические анализаторы и программы по созданию конкорданса, но и ряд «составных слов». Все тексты, включенные в материал для создания частотного словаря, членились на единицы автоматически и в ряде случаев полуавтоматически.

В основу анализа текстов, составляющих Национальный корпус калмыцкого языка, положена классификация единиц текста[2]. Поскольку в материал исследования входят и устные тексты, мы разделили весь массив слов на три части, которые отражают все особенности форм речи: речевые (вербальные), условно-речевые и неречевые (невербальные) — см. рис. 1.

Под речевыми понимаются вербальные единицы, обладающие обязательными признаками слова — фонетической, морфологической, лексико-семантической целостностью. Вербальные единицы текста включают в себя ряд классов.

1. Номинативный макрокласс: существительное, прилагательное, глагол (в том числе причастие и деепричастие), наречие, числительное. К этому классу относятся те слова, за которыми стоят понятия о предмете, о признаке, о действии (у них есть денотат). Их основной функцией является номинативная[3].

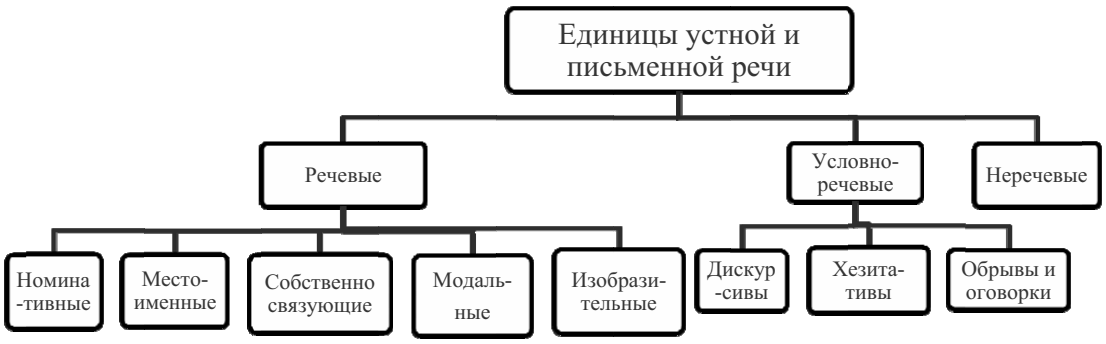
2. Местоименный макрокласс: местоимения-существительные, местоимения-прилагательные, местоимения-наречия, местоимения-числительные, местоимения-глаголы, выполняющие анафорическую и дейктическую функции в тексте.

[1] Многокомпонентные лексемы, «эквиваленты слова» — термин В. В. Виноградова. См., например: [Богданов, Рыжова 1997; Мустайоки, Копотев 2004; Венцов и др. 2004; Ягунова 2006; Крылов 2006; 2008].

[2] Классификация частично основана на концепции, разработанной группой авторов для составления семантического словаря (см. [Шведова 1998]).

[3] Следует отметить, что, помимо основных своих функций, те или иные речевые единицы могут выполнять и хезитационную функцию. Например, говорящий, затягивая артикуляцию гласных или согласных звуков, обдумывает свой следующий речевой фрагмент. Данная функция может быть актуальной не только для номинативного класса слов, но и для других единиц.

Рисунок 1
Единицы устной и письменной речи



3. Собственно связующий макрокласс: послелоги, союзы, связки и их аналоги. В эту группу входят слова, которые являются средствами синтаксической связи.

Эти три класса слов выполняют коммуникативно-информативную функцию на уровне текста, передавая фактуальную информацию; а также частично регулятивную функцию, т. е. оформляют взаимодействие говорящего и слушающего.

4. «Модальный» макрокласс: частицы, междометия, выражающие субъективное отношение и оценку говорящего.

5. К изобразительным словам относятся идеофоны, особый класс звукоподражательных и образных слов в калмыцком языке.

Поскольку в материал для анализа частотных характеристик современного калмыцкого языка включается и устная речь, то единицы, которые функционирует в разговорной речи и в то же время отражают особенности порождения устных текстов, рассматриваются нами в качестве единиц счета.

К условно-речевым единицам относятся дискурсивы, хезитативы различной структуры (в том числе слова-паразиты), обрывы, оговорки. Все они не имеют денотата и являются маркерами порождения высказывания [см. Леонтьев 1969: 133]. В группу вышеуказанных единиц входят следующие.

1. *Хезитативы* и дискурсивы, организующие дискурс. Они не несут в тексте смысловой нагрузки, но оформляют его структуру или заполняют паузы хезитации, могут иметь разную структуру (звук или слово).

Дискурсивы по своему характеру могут быть словами с размытой семантикой. Они появляются между более или менее законченными речевыми фрагментами — на границе высказываний, когда говорящий сопоставляет предыдущее высказывание согласно исходному замыслу и в то же время обдумывает уже следующую фразу. Возможно, эти единицы организуют композицию спонтанного текста. Это своего рода маркеры начала, продолжения и конца монолога. К тому же дискурсивы являются и сигналами для слушающего: «я начинаю, продолжаю или заканчиваю говорить».

К этой же группе относятся установочные дискурсивы, отражающие коммуникативные и психологические установки говорящего[1]. Они маркируют порождение более высокой единицы речи — текста.

Хезитативы возникают внутри высказываний. Это маркеры программирования – планирования, поиска, контроля[2].

2. Обрывы и оговорки, не обладающие фонетической целостностью. Они могут иметь «некоторый смысл» только в линейной последовательности высказывания.

К неречевым относятся паралингвистические элементы, такие как покашливание, смех, усмешки, вздохи, которые часто сопровождают речь говорящего. В естественных условиях появление таких элементов вполне закономерно. Круг неречевых элементов можно расширять до бесконечности, т. к. их источником может служить и сама ситуация общения, и адресат.

Следует оговорить, что в некоторых случаях паралингвистические элементы устного текста могут не только выражать эмоции говорящего, его субъективные состояния (в том числе физиологическое), но

[1] Об установочных маркерах см.: [Дараган 2000].

[2] Подробнее о программирующих маркерах см.: [Дараган 2000].

и являться одним из способов хезитации, например, кашель или вздох. О функции этих элементов можно судить на основании либо их уместности в контексте, либо их частотности появления в речи говорящего.

Ниже приведены правила анализа текстового материала в целях создания частотного словаря калмыцкого языка. Приведенные в настоящей работе правила основаны на принципах построения «Частотного словаря русского языка» [Частотный словарь … 1977] и «Грамматического словаря русского языка» [Зализняк 1987], которые переработаны применительно к калмыцкому языку, а также созданы для тех единиц, которых не существует в письменном литературном языке. Каждая часть речи имеет свой набор правил, который регламентирует работу морфологического анализатора в аспекте обработки линейной буквенной последовательности в элемент словника для частотного словаря.

НОМИНАТИВНЫЙ МАКРОКЛАСС

Имена существительные

1. Нарицательные имена

1.1. Исходной формой считается им. пад. ед. ч.:

- *модна* N.Gen=‘*дерево*’ → *модн* N.Nom=‘дерево’|Adj=‘*деревянный*’;
- *дегтрт* N.Dat=‘*книг*а’ → *дегтр* N.Nom=‘книга’;
- *эцкән* N.Gen.Refl=‘*отец*’→ *эцк* N.Nom=‘отец’.

1.2. Звательные формы существительных сводятся к им. пад. ед. ч.:

- Бадмаа N.Prop.Voc=‘Бадма-а’ → *Бадм* N.Prop.Nom=‘Бадма’;
- *аавааа* N.Prop.Voc=‘*дедушка-а*’ → *аав* N.Prop.Nom=‘дедушка’.

1.3. Супплетивные формы от разных основ считаются разными лексемами: *күн* N.Nom=‘человек’ ≠ *әмтн* N.Nom=‘люди’.

1.4. Собирательные существительные, употребляющиеся только во мн. ч., возводятся к им. пад.:

- малын N.Gen=‘скот’→ *мал* N.Nom=‘скот’;
- *турутна* N.Gen=‘копытные’ → *турутн* N.Nom=‘копытные’.

1.5. Сокращенные формы принимаются в качестве единицы анализа, хотя до сих пор еще не выявлен инвентарь сложившихся сокращений[1]. Буквенные аббревиатуры, которые являются собственными именами (*ХТ*, *СССР*, НКВД-д, *ГЭС*, *МТС* и т. п.), будут приведены в алфавитном и отдельном списках.

1.6. Сложные существительные с дефисом считаются одним словом: *хувцн-хунр* ‘одежда’, *ааһ-сав* ‘посуда’. Подобные последовательности не расчленяются на две единицы: *хувцн-хунр* ≠ *хувцн*, *хунр*, *ааһ-сав* ≠ *ааһ*, *сав*.

1.7. Существительные с послелогами даются как два отдельных слова:

- *модна деер* N.Gen=‘дерево’+Post=‘на’ → *модн* N.Nom=‘дерево’|Adj=‘*деревянный*’, *деер* Post=‘на’;
- *хотна ард* N.Gen=‘*хотон*’+Post=‘за’ → *хотн* N.Nom=‘хотон, село, поселок’, *ард* Post=‘за, позади’.

1.8. Фразеологизмы и устойчивые сочетания расчленяются на элементы, их образующие:

- *махлата мал* PhrC[2]=‘недотепа’ → *махла* N.Nom=‘шапка’, *мал* N.Nom=‘скот’;
- *хар күчн* PhrC=‘физическая сила’ → *хар* Adj=‘черный’, *күчн* N.Nom=‘сила’;
- *цаһан идән* PhrC=‘молочная пища’ → *цаһан* Adj=‘белый’, *идән* N.Nom=‘пища’.

Что касается элементов устойчивого выражения, которые не встречаются в свободном сочетании, заглавные формы выводятся искусственно.

2. Собственные имена

2.1. Собственные имена (*Дорж*, *Бадм*, *Баатр*, Чон) учитываются в качестве единицы счета. В ходе морфологического анализа и частичного снятия омонимии будет произведена дифференциация форм, являющихся нарицательными и собственными именами существительными, насколько это возможно. Дифференциальным признаком в снятии омонимии является написание токена с прописной буквы. В случаях, когда слово встречается в начале предложения (т. е. пишется с большой буквы), омонимия снималась вручную, где это было возможно. В случае невозможности снятия омонимии в корпусе в силу его большого объема та или иная единица будет учитываться и в группе нарицательных, и в группе собственных.

[1] См. подробно: [Куканова 2012в].

[2] Фразеологическая конструкция.

2.2. Если при собственных именах употреблены титулы или названия должностей, то они анализируются как отдельные слова:

- *Данзн нойн* N.Prop.Nom=‘*Данзан*’+N.Nom=‘нойн’ → *Данзн* N.Prop.Nom=‘Данзан’, *нойн* N.Nom=‘нойн’;
- Аюка хан N.Prop.Nom=‘Аюка’+N.Nom=‘хан’ → *Аюка* N.Prop.Nom=‘Аюка’, *хан* N.Nom=‘хан’.

2.3. Если названия чинов и должностей написаны с заглавной буквы, то они также учитываются в качестве самостоятельной единицы: Ohtphyn Дала Зая-пандита N.Gen=‘небо’ + Adv=‘*дала*’ + N.Prop.Nom=‘*Зая*’ + N.Nom=‘*пандита*’ → ohtphy N.Nom=‘’, дала Adv=‘’, Зая N.Prop.Nom=‘’, пандита N.Nom=‘’. Прозвища лиц, совпадающие с соответствующими нарицательными именами, отмечаются морфологическим анализатором как самостоятельные слова: *Мерган Баатр* N.Prop.Nom=‘*Мерген*’ + N.Nom=‘*баатр*’ → N.Prop.Nom=‘*Мерган*’, N.Nom=‘*баатр*’.

2.4. Географические названия (названия государств, стран, городов, рек, морей, озер, заливов и т. д.), названия планет, *месяцев, дней,* употребляющиеся как нарицательные, несмотря на то, что написаны со строчной буквы, не фиксируются (не снимается омонимия):

- *Лу җил* N.Prop.Nom=‘Лу’ + N.Dat = ‘год’ → *лу* N.Nom|Prop.Nom=‘дракон’[1], *җил* N.Dat=‘год’;
- *Алтн һасн* ‘*Полярная звезда*’ → *алтн* Adj=‘золотой’|N.Nom=‘*золото*’, *һасн* N.Nom=‘кол’.

2.5. Если в составе сложных имен собственных имеются компоненты, которые совпадают с нарицательными именами, то они возводятся к соответствующей начальной форме: *Хар Теңгст* N.Prop.Dat=‘Черное море’→ *хар* Adj=‘черный’, *теңгс* N.Nom=‘*море*’.

2.6. Названия статей, книг, изданий и организаций обрабатываются по тем же правилам: фиксируются лишь те компоненты, которые могут встретиться в качестве нарицательного имени, например: «*Хальмг Үнн*» N.Prop=‘Калмыцкая правда’ → хальмг Adj=‘калмыцкий’|N.Nom=‘калмык’, үнн N.Nom=‘правда’; «Тегин герл» N.Prop=‘Степной свет’ → тег N.Nom=‘степь’, герл N.Nom=‘свет’.

Имена прилагательные

1.1. В калмыцком языке имя прилагательное принадлежит к неизменяемому классу слов, находится в препозиции к определяемому слову и обозначает качество, признак, свойство предметов и явлений, например: *му* ‘плохой’, *сән* ‘хороший’, *ик* ‘большой’, *ахр* ‘короткий’ и т. д. Начальная форма прилагательных как класса неизменяемых слов совпадает со всеми формами.

1.2. Качественные прилагательные со значеним цвета, входящие в качестве компонента в сложное слово или в устойчивые сочетания, расчленяются на части: *цаһан седкл* Comp[2]=‘добродушие’ → *цаһан* Adj=‘белый’, *седкл* N.Nom=‘мысль; душа’; *улан хол* Comp=‘пищевод’ → *улан* Adj=‘красный’, *хол* N.Nom=‘горло’; *шар тосн* Comp=‘топленое масло’ → *шар* Adj=‘желтый’, *тосн* N.Nom=‘масло’.

1.3. В калмыцком языке отсутствуют формы сравнительной и превосходной степеней сравнения качественных имен прилагательных в калмыцком языке, однако имеется аналитический способ выражения значения интенсивности того или иного признака. Однако авторы «Грамматики калмыцкого языка» считают, что в калмыцком языке существуют способы выражения сравнительной и превосходной степени [Грамматика калмыцкого языка … 1983: 134]. На наш взгляд, в случае словосочетаний типа *салькнас хурдн* N.Abl=‘ветер’ + Adj=‘быстрый’ речь идет о сравнительной конструкции, а не способе выражения сравнения одного предмета с другим. Слова со словообразовательными суффиксами *-вр, -цр, -хн* (например, *улавр* Adj=‘красноватый’, *хатуцр* Adj=‘твердоватый’) также не являются способом выражения сравнительной степени, данные аффиксы придают мотивирующей основе значение интенсивности проявления того или иного признака, но никак не сравнения.

Конструкции со словами-интенсивами, придающие усиленное или ослабленное качество прилагательному, также расчленяются: *әвр күнд* Adv=‘очень’ + Adj=‘тяжелый’ → *әвр* Adv=‘очень’, *күнд* Adj=‘тяжелый’; *маш улан* Adv=‘*весьма*’ + Adj=‘*красный*’ → *маш* Adv=‘весьма’, *улан* Adj=‘красный’. Усилительная степень прилагательных образуется

[1] В данном случае анализатор выдает два морфологических разбора, поскольку не может снять омонимию форм имени собственного и нарицательного существительного.

[2] Компаунд.

с помощью полной или частичной редупликации, и такие конструкции расчленяются: *сәәхн-сәәхн цецгүд* Adj.Red[1]='красивый' + N.Pl.Nom='цветок' → *сәәхн* Adj='красивый', *сәәхн* Adj='красивый', *цецг* N.Nom='цветок'; *хурдн-хурдн мөрд* Adj. Red='быстр*ый'* + N.Nom='*лошадь'* → *хурдн* Adj='*быстрый*', *хурдн* Adj='быстрый', *мөрн* N.Nom='лошадь'; *хаб хар* Part.Emp='*очень*' + Adj='*черный*' → *хаб* Part.Emp='очень', *хар* Adj='черный'. Следовательно, при анализе не выделяются формы образования значения интенсивности признака, а указанные выше примеры идентифицируются как отдельные лексические единицы, обладающие своим собственным значением.

Глаголы

1. В качестве исходной формы было принято решение указывать форм**у** причастия в будущем времени на *-х*, согласно традиционной точке зрения.

2. Изъявительные, повелительные, желательные, предостерегательные формы глагола, атрибутивные (деепричастные и причастные) формы сводятся к форме на *-х*: *йовнавидн* V.Pres.1SPer='идти' → *йовх* V='идти'; *йовсн* Ptcpl.Pst='идти' → *йовх* V='идти'; *йовхар* Conv.Purp='идти' → *йовх* V='идти'; *йовдгнь* Ptcpl.Hab.3Poss='идти' → *йовх* V='идти'.

3. Видовые формы глагола также приводятся к исходной основе: *түркчкв* V.Compl.Pst='смазать; растирать' → *түркх* V='смазывать; растирать'; *көдлҗәлә* V.Dur2.Rem='**работать**' → *көдлх* V='работать'.

4. Залоговые формы глагола также приводились к исходной форме: *умшулна* V.Caus2.Pres='читать' → *умшх* V='читать'; *орлцҗ* Conv.Soc.Ipfv='входить'→ *орх*; *эвцлдхлә* Conv.Recp.Cond='соглашаться' → *эвцх* V='соглашаться'; *даалһгдсн* Ptcpl. Caus1.Pass.Pst='ручаться; терпеть; резать' → *даах* V='ручаться; терпеть; резать'.

5. В составном глагольном сказуемом (так называемых сложных глаголах[2]) все компоненты возводятся к инфинитиву: *авад оркна* Conv.Ant='брать' + V.Pst='ставить; складывать' → *авх* V='брать', *оркх* V='ставить; складывать'; *бууһад ирҗәнә* Conv.Ant='*спускаться*' + V.Dur2.Pst= '*прийти*'→ *буух* V='*спускаться*', *ирх* V='*прийти*'; *гүүлдҗ ирлдв* Conv.Recp. Ipfv='бежать' + V.Recp.Pst='прийти'→ *гүүх* V='*бежать*', *ирх* V='*прийти*'.

6. Устойчивые глагольные сочетания расчленяются на составляющие их лексемы: *хар гөрлә харһх* PhrC='*быть ложно подозреваемым*' → *хар* Adj='черный', *гөр* N.Com='подозрение', *харһх* V='встречаться; сходиться'; *улан махн болтлнь цокх* PhrC='жестоко избивать кого-либо полусмерти' → *улан* Adj='красный', *махн* N.Nom='мясо', *болх* V='становиться', *цокх* V='бить'.

Наречия

Наречия в калмыцком языке — это неизменяемая часть речи, обозначающая признак действия и признак качества. Важнейшим морфологическим признаком наречий является их соотносительность с именными частями речи, глаголами и отглагольными формами. В этой поздней по своему происхождению частью речи ее количественный состав постоянно растет за счет адвербализованных форм существительных, местоимений, деепричастий и других разрядов, что существенно затрудняет снятие омонимии [Грамматика калмыцкого языка 1983: 259].

1. Некоторые адвербиализированные формы существительных, местоимений, деепричастий и других разрядов можно отличить от соответствующих омонимичных употреблений слов по некоторым формальным показателям, хотя выявление этого отличия — задача уже другого исследования. Например, имена существительные в орудном падеже зачатую переходят в разряд наречий, однако при наличии возвратных и притяжательных частиц они рассматриваются как существительные: *күчәр кеҗәнә* Adv='*сильно*' + V.Dur2.Pst='*делать*'→ *күчн* N.Nom='*сила*', *кех* V='*делать*'; но *әмн күчәрн* хәәкрәд Adj='жизненный' + N.Instr. Refl='сила' + Conv.Ant='кричать' → *әмн* N.Nom='жизнь', *күчн* N.Nom='сила', *хәәкрх* V='кричать'. Существительные в исходном падеже могут также переходить в разряд наречий, однако при присоединении частиц притяжания остаются в разряде существительных: *хаҗуһас соңссн* Adv='сбоку, со стороны, извне' + Ptcpl.Pst='слушать' →*хаҗуһас* Adv='сбоку, со стороны, извне', соңсх V='слушать'; но *хәврһәснь хәләсн* N.Abl.3Poss='бок; сторона' + Ptcpl. Pst='смотреть' → *хәврһ, хәләх.*

2. В некоторых случаях наречия и падежные формы существительных формаль-

[1] Red — редупликация.

[2] См. подробно [Баранова 2009].

но не дифференцированы друг от друга, различить их можно только в контексте. Здесь наблюдается грамматическая омонимия, при которой имена существительные в косвенных падежах, с одной стороны, и наречия, с другой, выполняя различные функции в предложении, несут в себе различное функциональное содержание, внешне оставаясь идентичными друг другу. При автоматической обработке больших объемов текста разграничение таких грамматических омонимов в полной мере не представляется возможным.

3. Некоторые имена существительные в дательном и соединительном падежах, несущие временные и пространственные значения, принимая посессивные и рефлексивные частицы, считаются адвербиализованными, поэтому записываются как отдельные слова: намртнь Adv=‘осенью’ → *намртнь* Adv=‘осенью’; *цаглань* Adv=‘своевременно’ → *цаглань* Adv=‘своевременно’.

4. Предельное деепричастие на *-тл* при присоединении возвратной частицы *-ан* (*-ән*) переходит в разряд наречий и при разборе учитывается как отдельное слово: *үктлән эн һос эдлҗ өмсхв* Adv=‘до смерти’ + Pron.Dem=‘этот’ + N.Acc=‘сапог’ + Conv.Ipfv=‘пользоваться’ + Ptcpl. Fut.1SPer=‘надевать’→ *үктлән, эн, һосн, эдлх, өмсх*.

5. Большую группу наречий, несущих временное значение, составляют сложные сочетания соединительного деепричастия на *-җ* и вспомогательного глагола *бәәх*. Например, *тиигҗәтл* ← *тиигҗ* + *бәәтл; иигҗәһәд* ← *иигҗ* + *бәәһәд*. При разборе эти формы не разбирались как два отдельных слова, а анализировались как одно: *тиигҗәтл* → *тиигҗәтл; иигҗәһәд* → *иигҗәһәд.*

6. Фразеологизированные наречные сочетания и наречные конструкции расчленяются: *һарх зуур* PhrC=‘при выходе’ → *һарх* V=‘выходить’, *зуур* Adv=‘перед, когда’; *гем уга* PhrC=‘болезней (?) нет’→ *гем* N.Nom=‘[1]*болезнь;* [2]вина*’*, *уга* Part.Neg=‘не’.

7. Словообразовательные и орфографические варианты наречий записываются как отдельные слова: *деер* и *деерәкшән; дор* и *дорагшан*; *деерәкшән и деегшән*; *өмәрән* и *өмнәгшән*.

МЕСТОИМЕННЫЙ МАКРОКЛАСС

Местоимения

1. Исходные формы личных местоимений (1 и 2-е лица), изменяющихся по падежам и числам и имеющих супплетивные основы, приводятся к форме именительного падежа соответствующего числа:

- *мини* Pron.Pers.Gen=‘я’→ *би* Pron.Pers. Nom=‘я’;
- *нанас* Pron.Pers.Abl=‘я’ → *би* Pron.Pers. Nom=‘я’;
- *чини* Pron.Pers.Gen=‘ты’ → *чи* Pron.Pers. Nom=‘ты’;
- *чамд* Pron.Pers.Dat=‘ты’ → *чи* Pron.Pers. Nom=‘ты’;
- *тадн* Pron.Pers.Nom=‘Вы’ → *тадн* Pron. Pers.Nom=‘Вы’;
- *танар* Pron.Pers.Instr=‘Вы’→ *та* Pron. Pers.Nom=‘Вы’.

Диалектные и другие варианты личных местоимений, имеющие отклонения в формах падежа, также приводятся к исходной форме: *намла* Pron.Pers.Com.Dialmorph=‘я’ → *би* Pron.Pers.Nom=‘я’; *намас* Pron.Pers. Abl.Dialmorph =‘я’ → *би* Pron.Pers.Nom=‘я’.

2. Эксклюзивная (*бидн*) и инклюзивная (*мадн*) формы множественного числа местоимения 1-го лица приводятся к своей исходной форме:

- *мана* Pron.Pers.Gen=‘мы’ → *бидн* Pron. Pers.Nom=‘мы’;
- *манар* Pron.Pers.Instr=‘мы’ → *бидн* Pron. Pers.Nom=‘мы’;
- *маднта* Pron.Pers.Assoc=‘мы’ → *мадн* Pron.Pers.Nom=‘мы’;
- *маднур* Pron.Pers.Dir=‘мы’ → *мадн* Pron. Pers.Nom=‘мы’.

3. В калмыцком языке предметно-указательные местоимения *эн, эдн, тер, тедн* также используются для обозначения 3-го лица, вследствие чего появляются омонимичные разборы (Pron.Pers|Dem). В косвенных падежах местоимения *эн*, *тер* имеют вариативные формы-основы — *энүнә, үүнә; терүнә, түүнә* (OrphV[1]), исходной формой в данном случае считается форма именительного падежа соответствующего числа:

- *энүнә* Pron.Dem.Gen.OrphV=‘этот’ → *эн* Pron.Dem.Nom=‘этот’;
- *үүнә* Pron.Dem.Gen.OrphV=‘этот’ → *эн* Pron.Dem.Nom=‘этот’;

[1] OrphV — орфографический вариант.

- *терүнд* Pron.Dem.Dat.OrphV=‘тот’→ *тер* Pron.Dem.Nom=‘тот’;
- *түүнә* Pron.Dem.Gen.OrphV=‘тот’ → *тер* Pron.Dem.Nom=‘тот’.

4. Косвенные падежные формы качественно-указательных (*иим, тиим*) и количественно-указательных (*эдү, тедү*) местоимений приводятся к именительному падежу: *иимин* Pron.Dem.Acc=‘такой’ → *иим* Pron.Dem.Nom=‘такой’; *эдүд* Pron. Dem.Dat=‘столько’ → *эдү* Pron.Dem. Nom=‘столько’.

5. Пространственно-указательные местоимения *энд*, *тенд*, *эдүкнд*, *тедүкнд*, а также глагольно-указательные *иигх*, *тиигх* принимались за местоименные слова и, соответственно, разбирались как наречия и глаголы.

6. Определенные местоимения *цуг*, *цуһар*, *цугтан*, изменяющиеся по падежам, принимались за одно слово и возводились к исходной форме *цуг*, *цуһар*, *цугтан*, соответственно:

- *цугиг* Pron.Qua.Acc=‘весь’→ *цуг* Pron. Qua.Nom=‘весь’;
- *цуһаратань* Pron.Qua.Assoc.3Poss= ‘все’→ *цуһар* Pron.Qua.Acc=‘все’;
- *цугтаһарнь* Pron.Qua.Instr.3Poss =‘вместе’ → *цугтан* Pron.Qua.Acc=‘вместе’.

7. Косвенные формы определительных местоимений *хамг*, *бүгд*, *зәрм* возводились к исходной форме именительного падежа:

- *хамгиг* Pron.Qua.Acc=‘все’ → *хамг* Pron. Qua.Nom=‘все’;
- *зәрмлә* Pron.Qua.Com=‘некоторый’ → *зәрм* Pron.Qua.Nom=‘некоторый’;
- *бүгдәс* Pron.Qua.Abl=‘все’ → *бүгд* Pron. Qua.Nom=‘все’.

Необходимо отметить, что местоимение *зәрм* в форме дательного падежа и притяжательной частицей *-ән* переходит в разряд местоименных наречий со значением времени. В этом случае оно рассматривается как отдельное слово *зәрмдән*.

8. Возвратные местоимения *эврән*, *бий*, склоняющиеся по обычному типу и способные наращивать частицы притяжания, возводились к исходной форме: *эврәг* Pron.Refl. Acc=‘сам’ → *эврән* Pron.Refl.Nom=‘*сам*’, *бийәр* Pron.Refl.Instr=‘сам’ → *бий* Pron.Refl. Nom=‘*сам*’. Множественное число возвратного местоимения *бийснь* также возводится к форме именительного падежа в единственном числе (*бий*).

9. Вопросительные местоимения в калмыцком языке, в зависимости от семантической нагрузки, разбиваются на несколько групп.

9.1. Предметно-вопросительные местоимения *кен*, *юн* имеют полную парадигму склонения и могут присоединять частицы усиления *-чн*, в результате автоматического анализа приводятся к именительному падежу:

- *кенд* Pron.Inter.Dat=‘кто’ → *кен* Pron. Inter.Nom=‘что’;
- *кениг* Pron.Inter.Acc=‘кто’ →*кен* Pron. Inter.Nom=‘что’;
- *юута* Pron.Inter.Assoc=‘что’ → *юн* Pron. Inter.Nom=‘*ч*то’;
- *юнасчн* Pron.Inter.Abl.EmpPart=‘что’ → *юн* Pron.Inter.Nom=‘что’.

Конструкции с предметно-вопросительными местоимениями с усилительными словами *чиг*н, *болвчн* расчленяются: *юн чиг*н Pron.Qua.Nom=‘всякий’ → *юн* Pron.Inter.Nom=‘что’, *чиг*н Part.Emp; *кениг чиг*н Pron.Qua.Acc=‘всякий’→ *кен* Pron. Inter.Nom=‘кто’, *чиг*н Part.Emp; *кен болвчн* Pron.Qua.Nom=‘хоть кто’→ *кен* Pron.Inter. Nom=‘кто’, *болвчн* Conj=‘*болвчн*’.

9.2. Качественно-вопросительное местоимение *ямаран* при автоматическом анализе приводится к форме именительного падежа: *ямаранд* Pron.Inter.Dat=‘какой’ → *ямаран* Pron.Inter.Nom=‘какой’; *ямараниг* Pron.Inter.Acc=‘какой’ → *ямаран* Pron.Inter. Nom=‘какой’.

9.3. Качественно-вопросительное местоимение *кедүдгч*, образованное от количественно-вопросительного местоимения *кедү*, считается отдельным словом.

9.4. Количественно-вопросительное местоимение *кедү*, имеющее полную парадигму склонения, приводится к форме именительного падежа: *кедүһәр* Pron. Inter.Inst=‘сколько’ → *кедү* Pron.Inter. Nom=‘сколько’.

9.5. Пространственно-вопросительные местоимения[1] *аль*, *альд*, *хама*, имеющие неполную парадигму склонения, возводятся к именительному падежу*:* *альдас* Pron.Inter. Acc=‘откуда’ → *аль* Pron.Inter.Nom=‘где’, *хамаһур* Pron.Inter.Dir=‘куда’ → *хама* Pron. Inter.Nom=‘где’.

[1] Нередко эти местоимения в форме дательно-местного падежа с частицами притяжания (или без них) могут переходить в разряд местоименных наречий.

9.6. Вопросительно-временное местоимение *кезә* не имеет полной парадигмы склонения. Косвенные формы этого местоимения *кезәнә*, *кезәһә*, *кезәнәс*, *кезәһәс* перешли в разряд местоименных наречий и принимаются за отдельные слова. Тоже самое и с глагольно-вопросительным местоимением *яах*, которое обладает всеми свойствами глагола, и рассматривается нами как отдельное слово.

10. Неопределенные местоимения в калмыцком языке образованы сочетанием различных именных частей речи, поэтому при разборе расчленяются на составные части: *нег гер* Pron.Ind=‘некий’|Num.Card=‘один’ + N.Nom=‘дом’ → *нег* Pron.Ind=‘некий’|Num.Card=‘один’, *гер* N.Nom=‘дом’. В случае с нег появляется омонимия формы неопределенного местоимения и числительного, которую снимать придется вручную.

11. Отрицательные местоимения, образованные аналитическим образом (путем присоединения отрицательных слов *биш*, *уга*), разделяются на составные части: *юн чиг уга* Pron.Neg=‘ничто’→ *юн* Pron.Inter=‘что’, *чиг* Part.Emp, *уга* Part.Neg=‘нет’; но *ямаранчн биш* Pron.Neg=‘*н*икакой’ → *ямаран* Pron.Inter=‘какой’, *биш* Part.Neg=‘нет’.

12. Возвратное местоимение *эврән* в орудном падеже при наращивании частицы притяжания выступает в роли наречия, поэтому анализируется как отдельное слово: *эврәһәрн* Adv → *эврәһәрн* Adv.

13. Именные местоимения в форме двойного склонения родительного и орудного падежей (местоимение + *ин* + *әр*) также переходят в разряд наречий, поэтому учитываются как отдельные слова: *миниһәр* Adv → *миниһәр* Adv, *чиниһәр* Adv → *чиниһәр* Adv; *танаһар* Adv → *танаһар* Adv.

СОБСТВЕННО СВЯЗУЮЩИЙ МАКРОКЛАСС

Послелоги

1. Послелоги на данном этапе (т. е. на этапе морфологического анализа без снятия омонимии) не отграничиваются от омонимичных употреблений именных слов, наречий, глагольных форм и т. д. В калмыцком языке послелоги, в отличие от именных слов, наречий, глагольных форм, не являются самостоятельными лексическими единицами с присущими им морфологическими, синтаксическими и семантическими признаками, выступая в предложении в качестве уточнителей значений при именных частях речи. В предложении послелоги сочетаются с именными словами в строгой последовательности — всегда в постпозиции, тем самым отличаясь, в первую очередь, от наречий и других частей речи: *герин ца* N.Gen=‘дом’ + Post=‘за’ → *гер* N.Nom=‘дом’, *ца* Post=‘за’; *үдин алднд* N.Gen=‘полдень’ + Post=‘около; почти’→ *үд* N.Nom=‘полдень’, *алднд* Post=‘около; почти’.

Союзы

1. Союзы в калмыцком языке по своему происхождению, составу и семантике значительно отличаются друг от друга. Значительная их часть образована от глагольных форм, а также наречий, местоимений, частиц и послелогов. По этой причине при автоматическом морфологическом разборе в ряде случаев наблюдается грамматическая омонимия, при которой союзы и образовавшие их знаменательные или служебные слова внешне совпадают. При подсчете частоты того или иного слова в целях создания более объективной картины структурно-вероятностной системы языка множественные формы будут учитываться в каждой группе.

Например, токены, образованные от глагола *гих*, из-за возможности совпадения с атрибутивными (причастными и деепричастными) формами и союзами разбираются морфологическим анализатором и как союз, и как глагол. В граммеме дается: *гиһәд* Post|Conv → *гих* V=‘говорить’|*гиһәд* Post; *гиҗ* Post|Conv → *гих* V=‘говорить’|*гиҗ* Post.

2. Сложные и составные союзы расчленяются на отдельные лексемы: *тегәд чиг* Conj=‘поэтому, потому’ → *тегәд* Pron=‘затем’, *чиг* Part.Emp; *хәрнь зуг* Conj → *хәрнь* Conj=‘однако’, *зуг* Conj=‘но’; *бас чиг* Conj → *бас* N=‘[1]*баз*; [2]*бас*’|Conj=‘тоже’|Adv=‘снова, опять’, *чиг* Part.Emp и т. п.

3. Двойные союзы фиксировались как разные словоупотребления: *аль,* < ...> *аль* Conj=‘или … или’ → *аль* Conj=‘или’, *аль* Conj=‘или’; *эс гиҗ* <...> *эс гиҗ* → *эс* Part.Neg=‘не’, *гиҗ* Post, *эс* Part.Neg=‘не’, *гиҗ* Post.

«МОДАЛЬНЫЙ» МАКРОКЛАСС

Частицы

1. Частицы в калмыцком языке в большинстве своем примыкают к слову в силу сильной редукции, построить частоты здесь достаточно трудоемко. В записях граммем частицы фиксируются с указанием их разряда. Вследствие этого мы решили зафиксировать частоты только тех частиц (частицы отрицания *эс*, запрета *бичә* (*бичкә*) и т. д.), которые пишутся отдельно от слова — т. е. через пробел. Например, *келсн угав* Ptcpl. Pst=‘говорить’ + Part.Neg.1SPer=‘нет’ → *келх* V=‘ говорить’, *уга* Part.Neg=‘нет’; *бичә йов* Part.Neg=‘*нет*’ + V.Impr.2SPer=‘*идти*’ → *бичә* Part.Neg=‘*нет*’, *йовх* V=‘*идти*’.

2. Список частиц как морфемных элементов слова генерируется отдельным списком с указанием их разряда.

Междометия

1. Междометия с дефисом, функционально отличающиеся от соответствующих бездефисных форм, записываются как отдельное слово: *чаг-чагра* Intj → *чаг-чагра* Intj; *на-ца* Intj → *на-ца* Intj.

2. Фонетически удлиненные междометия даются как одно слово: *а-а-а* Intj → *а* Intj; *о-ой* Intj → *ой* Intj.

3. Составные (производные) междометия расчленяются на отдельные лексемы, если включают нетождественные компоненты: *чиш тәтә* Intj → *чиш* Intj, *тәтә* Intj; *ю тәтә* Intj → *ю* Intj, *тәтә* Intj; *дәрк әәлдг* Intj → *дәрк* Intj, *әәлдг* Intj.

4. Составные (производные) междометия расчленяются на отдельные лексемы, если включают тождественные компоненты: *түш-түш* Intj → *түш* Intj, *түш* Intj; *тип-тип* Intj → *тип* Intj, *тип* Intj.

Идеофоны

1. В калмыцком языке есть изобразительные слова, которые обладают самостоятельным значением и особой структурой [Грамматика калмыцкого языка 1983: 303]. При разборе такие слова расчленяются на составляющие: *пард гих* Idf → *пард* Idf, *гих* V=‘говорить’, *таш гих* Idf → *таш* Idf, *гих* V=‘говорить’, *гилс гих* Idf → *гилс* Idf, *гих* V=‘говорить’.

2. Редупликативные формы рачленяются на составляющие элементы и считаются отдельно: *бүр-бүр гих* Idf → *бүр* Idf, *бүр* Idf, *гих* V=‘говорить’.

Условно-речевые единицы

1. Паузы хезитации, заполненные неречевыми звуками, фиксируются как отдельные единицы: *э-э*, *э-м*, *м-м*, *а-а*, *а-м*.

2. Словесные заполнители пауз хезитации (так называемые *слова-паразиты*) также анализируются как одна единица, поскольку они, во-первых, не имеют собственного лексического значения, во-вторых, выполняют функцию хезитации (обдумывания), в-третьих, как правило, реализуются в звучащей речи как одно фонетическое слово.

3. Паралингвистические элементы также даются как отдельные единицы, поскольку они выполняют, как правило, хезитационную или эмотивную функции. Например, <*смех*>, <*вздох*>, <*кашель*>и т. д.

4. Обрывы слов фиксируются как отдельные словарные единицы: *на...*, *өр…*, *у…*, *ба...*, *ке...*, *ю...* и т. п.

Таким образом, приведенные правила анализа текстового материала могут служить эскизом анализирующей модели переработки сегментов текста в элементы словаря. Калмыцкий язык как язык с богатым словоизменением создает определенные трудности для компилирования частотного словаря, так как многие словоформы в текстах омонимичны (ср. словоформу *көвәд* как формы от омонимичных глаголов *көвх*[1, 2] и существительного *көвә*, словоформу *үүлд*, представляющую леммы *үүлн* и *үүл*, слова типа *нарн* и *Нарн*). Тем не менее, в частотном словаре исходная форма слова, или лемма, должна быть приписана любой словоформе однозначно, чтобы программа могла однозначно посчитать частоты того или иного сегмента текста.

Это даст возможность в будущем собрать данные о частотности не только отдельных словоформ, но и лексем, а также об употребительности тех или иных грамматических категорий. На основе же метатекстовой информации будут сравниваться между собой частотные списки на отдельных выборках корпуса (по функциональным стилям, по времени создания текста). Во второй части данной работы, подготовленной к публикации в следующем номере журнала, будут приведены «верхушки» частотных списков словоформ, слов, слов внутри одной части речи, граммем.

Литература

Баранова В. В. Сложные глаголы в калмыцком языке // Исследования по грамматике калмыцкого языка/ ред. С. С. Сай, В. В. Баранова, Н. В. Сердобольская. СПб.: Наука, 2009. Том V. Ч. 2). С. 255–310. (ACTA LINGUISTICA PETROPOLITANA. Труды Ин-та лингвист. исслед. РАН).

Бертагаев Т. А. Синтаксис современного монгольского языка в сравнительном освещении. Простое предложение. М.: Наука, 1964. 300 с.

Богданов С. И., Рыжова Ю. В. Русская служебная лексика. Сводные таблицы. СПб.: изд-во СПб. ун-та, 1997. 293 с.

Венцов А. В., Грудева Е. В., Касевич В. Б., Ягунова Е. В. Об идиомах в Национальном корпусе русского литературного языка // Компьютерная лингвистика-2004. Тезисы международной конференции. 12–14 октября 2004 г. СПб., 2004. С. 17–18.

Гак В. Г. Слово // Лингвистический энциклопедический словарь / под ред. В. Н. Ярцевой. М.: Советская энциклопедия, 1990 [электронный ресурс] // URL: http://tapemark.narod.ru/les/464c.html (дата обращения: 07.03.2012).

Грамматика калмыцкого языка: фонетика и морфология. Элиста: Калм. кн. изд-во, 1983. 336 с.

Дараган Ю. В. Функции слов-«паразитов» в русской спонтанной речи [электронный ресурс]// URL: http://www.dialog-21.ru/materials/archive.asp?id=6260&vol=6077&y=2000. 2000 (18.05.2008).

Долинский В. А. Квантитативная лингвистика в исследовании текста // Алфавит: Строение повествовательного текста. Синтагматика. Парадигматика. Смоленск: СГПУ, 2004. С. 283–324.

Зализняк А. А. Грамматический словарь русского языка: Словоизменение: Около 100 000 слов. 3-е изд., стереотип. М.: Рус. яз., 1987. 880 с.

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

Касевич В. Б. Элементы общей лингвистики. М.: Наука, ГРВЛ, 1977. 177 с.

Копотев М. Несмотря на, потому что, или многокомпонентные единицы в аннотированном корпусе русских текстов // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог'2004» («Верхневолжский», 2–7 июня 2004 г.). М., 2004. (URL: http://www.dialog-21.ru/Archive/2004/Kopotev.htm (17.07.2008)).

Крылов С. А. Измерение частотности синтаксических молекул (на материале генерального корпуса русского языка) // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» («Бекасово», 48 июня 2008 г.). Вып. 7 (14). М., 2008. С. 254–261.

Крылов С. А. Об инвентарных и конструктивных единицах языка // Язык и речевая деятельность. 2003. Вып. 6. СПб., 2006. С. 9–26.

Крылов А. С. Опыт изучения современного монгольского языка в количественном аспекте // Вопросы языкознания. 2013. № 5. С. 46–58.

Куканова В. В. О корпусе калмыцких текстов: краткий обзор проблем графематического анализа // Научное наследие проф. А. Ш. Кичикова и актуальные проблемы современной калмыцкой филологии и культуры (Кичиковские чтения). Материалы Региональной научной конференции, посвященной 90-летию со дня рождения профессора А. Ш. Кичикова (21 декабря 2011 г., Элиста). Элиста: Изд-во Калм. гос. ун-та, 2012в. С. 61–63.

Куканова В. В. Словоизменительные типы в калмыцком языке в свете автоматической обработки текстов (на примере имени существительного) // Вестник Калмыцкого института гуманитарных исследований РАН. 2012а. № 2. С. 168–177.

Куканова В. В. Словоизменительные типы в калмыцком языке в свете автоматической обработки текстов (на примере имени существительного) – II // Вестник Калмыцкого института гуманитарных исследований РАН. 2012б. № 3. С. 151–161.

Куканова В. В., Бембеев Е. В., Мулаева Н. М., Очирова Н. Ч. Метаразметка в Национальном корпусе калмыцкого языка // Вестник Калмыцкого государственного университета. 2012а. № 3. С. 67–72.

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. 2012б. № 3. С. 138–150.

Лёнгрен Л. (ред.). Частотный словарь современного русского языка [Lönngren, Lennart. The Frequency Dictionary of Modern Russian. Acta Univ. Ups., StudiaSlavicaUpsaliensia Uppsala 32]. Uppsala, 1993. 188 с.

Леонтьев А. А. Психолингвистические единицы и порождение речевого высказывания. М.: Наука, 1969. 307 с.

Ляшевская О. Н. , Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. 1090 с.

Мустайоки А., Копотев М. К вопросу о статусе эквивалентов слова типа потому что, в зависисмости от, к сожаленью // Вопросы языкознания. М., 2004. № 3. С. 88–107.

Степанова Е.М. Частотный словарь общенаучной лексики. М.: Просвещение, 1976. 87 с.

Частотный словарь русского языка / под ред. Л. Н. Засориной. М.: Русский язык, 1977. 936 с.

Шведова — Русский семантический словарь. Толковый словарь, систематизированный по классам слов и значений / Ин-т рус. яз. им. В. В. Виноградова РАН; под общ. ред. Н. Ю. Шведовой. Т. 1. М.: Азбуковник, 1998. XXV+807 с. Т. 2. М.: Азбуковник, 2000. XXXII+762 с. Т. 3. М.: Азбуковник, 2003. 720 с. Т. 4. М.: ИРЯ РАН, 2007. 952 с.

Ягунова Е. В. Неоднословные целостности в словаре и в корпусе // Корпусная лингвистика-2006. Труды международной конференции. 10–14 октября 2006. СПб., 2006. С. 395–412.