4. Мянгад Эрдэмт. Равджамба Зая-Пандида Намхайджамц судлал. Biblioteka oiratika Volium 7. – Улаанбаатар. 2008. С. 112.
5. Александр Зорин. Гимны Таре. С. 26.
6. Рерих Ю.Н. Тибетско-русско-английски словарь с санскритскими параллелями. – М.: Изд-во «Наука», 1983.
7. Amitabhayin magtāl kemēkü orošiboi. Текст из коллекции автора.
8. Торчинов Е. А. Введение в буддизм. – СПб.: Амфора, 2005. С.17
9. Рерих Ю.Н. Тибетско-русско-английски словарь с санскритскими параллелями. – М.: Изд-во «Наука», 1983.
10. Далай – Лама. Сутра сердца. – М.: Изд-во «Океан Мудрости», 2008. С. 112.
11. Рерих Ю.Н. Тибетско-русско-английски словарь с санскритскими параллелями. – М.: Изд-во «Наука», 1983.
12. Сухбаатар О. Монгол хэлний харь угийн толь. Монгол улсын шинжлэх ухааны академии. Хэл зохиолын хүрээлэн. – Улаанбаатар, 1997.
13. Сухбаатар О. Монгол хэлний харь угийн толь. Монгол улсын шинжлэх ухааны академии. Хэл зохиолын хүрээлэн. – Улаанбаатар. 1997.
14. Gdugs-dkar-mchog-grub-ma-bzhugs-so. Сутра превосходнейшей практики богини белого зонта. Текст из коллекции автора.
15. Tögünčilen bolugsani usnir-ēce γarugsan caγān šükürtü busad-tu ülü ilaγugdagči yeke xariu xariulugči sayitur bütegsen kemēkü togtōl. Текст из коллекции автора.
16. Gdugs-dkar-mchog-grub-ma-bzhugs-so. Сутра превосходнейшей практики богини белого зонта. Текст из коллекции автора.
17. Tögünčilen bolugsani usnir-ēce γarugsan caγān šükürtü busad-tu ülü ilaγugdagči yeke xariu xariulugči sayitur bütegsen kemēkü togtōl. Текст из коллекции автора.
18. Позднеев. А.М. Очерки быта буддийских монастырей и буддийского духовенства в Монголии в связи с отношениями сего последнего к народу. Репринтное издание. – Элиста: Калм. кн. изд-во, 1993. С. 209
19. А. Зорин. Гимны Таре. – М.: Открытый Мир, 2009. С. 28.

**Список принятых сокращений**

Тиб. – тибетский
Ойр. – ойратский
Санскр. – санскрит
Русс. – русский

***

***В.В. Куканова,***
*КИГИ РАН*

## О КОРПУСЕ КАЛМЫЦКИХ ТЕКСТОВ: КРАТКИЙ ОБЗОР ПРОБЛЕМ ГРАФЕМАТИЧЕСКОГО АНАЛИЗА*

**Статья подготовлена при поддержке проекта «Национальный корпус калмыцкого языка» подпрограммы фундаментальных исследований Президиума РАН «Создание и развитие корпусных ресурсов по языкам народов России» программы «Корпусная лингвистика» (2012–2014) и проекта РГНФ «Национальный корпус калмыцкого языка» (12-04-12047, тип «в»)*

Корпусные базы данных создаются во многих странах мира, однако их объектом выступали, с одной стороны, европейские, с другой – международные языки, как, например, английский, китайский, русский и др. Многие проблемы, связанные с представительностью, токенизацией, лемматизацией и снятием омонимии, уже решены при разработке корпусов языков, которые в основном являются флективными. При создании подобных систем для малых языков лингвисты сталкиваются с проблемами совсем иного рода. В случае создания текстовых баз данных для европейских или международных языков лингвист имеет дело с уже относительно устоявшейся системой орфографии, пунктуации, грамматики, функциональных сфер использования языка и т. д. по сравнению с исчезающими языками, развитие которых не столь устойчивое и стабильное, как в широко распространенных языках. Но при всех плюсах последних, естественно, существуют иные проблемы, как, например, унификация различных классификаций текстов, поскольку последних бессчетное множество и «подогнать» их под одни рамки практически невозможно. Лингвист, создавая информационно-справочные системы для языка малого народа, сталкивается с проблемами, например, репрезентативности материала в корпусе. Последняя заключается не в том, как охватить все многообразие текстов, а

в том, чтобы представить все то, что существует на данный момент, в том объеме и с теми диспропорциями, которые имеют место быть в реальной языковой ситуации. Проблемы графики и орфографии усложняют процесс автоматической обработки текстов на языках малых народов, поскольку правила обозначения некоторых звуков до сих пор еще не выработаны окончательно (например, в калмыцком языке так называемые «неясные» гласные).

Работы по созданию корпусов по малым языкам ведутся в некоторых (волонтерских или государственных) организациях, однако их число на фоне количества вымирающих языков совсем не значительно. Тем не менее даже и эти проекты – это большой вклад в дело сохранения этнокультурного наследия народов, находящихся на грани исчезновения. Например, в Дагестанском государственном университете проводятся работы по оцифровке текстов и разработке программных модулей для автоматической обработки, а также в Институте монголоведения и тибетологии РАН создают электронный корпус бурятского языка [1]. Проект по созданию Национального корпуса калмыцкого языка ведется и в Калмыцком институте гуманитарных исследований РАН и направлен, прежде всего, на решение проблемы сохранения, развития и исследования калмыцкого языка, который является достоянием не только российской, но и мировой культуры. На данный момент собрана представительная коллекция текстов как на кириллице, так и «тодо бичиг» объемом в 4 млн словоупотреблений, создан обратный, или грамматический, словарь калмыцкого языка, необходимый для создания морфологического анализатора, разработана архитектура метаописания текстов [2].

В данной статье мы попытаемся не только описать некоторые (насколько позволит ограниченный объем статьи) проблемы графематического анализа в аспекте создания Национального корпуса калмыцкого языка, но и поставить вопрос о лингвистическом статусе многокомпонентных единиц (так называемых эквивалентов слова – термин В. В. Виноградова, 1947) и возможности их описания в лексикографическом и грамматическом планах в калмыцком языкознании.

Прежде чем приступить к выполнению задачи, объясним, что предполагает этот процесс. Графематический анализ, или токенизация, – это автоматическая обработка, которая заключается в разбиении линейного потока символов на конструктивные языковые единицы (словоформу, предложение) и более крупные (абзац, главу, заголовок) от предыдущего и последующего элементов текста и т. д. [3; 106]. Как видим, это самый простой вид анализа текстового материала, но от которого зависит правильность работы морфологического и синтаксического парсеров. На этом этапе осуществляется также и предморфологический анализ текстов – объединение в одно целое и разбиение на составляющие.

Сегментация на предложения в калмыцком языке совпадает с правилами, выработанными для русского языка (см. [4]). В качестве делиметра (разделителя) для сегментации линейного потока символов на калмыцком языке на словоформы служит пробел между словами (так называемый whitespace) так же, как и во многих других языках (английском, русском). Однако, как и в русском языке, это порождает ряд проблем в обработке линейного потока.

Приведем некоторые из них.

1. Аббревиатуры, употребление которых в калмыцком языке началось только в период советской власти. К тому же сложносокращенные слова являются заимствованными словами и передаются в калмыцком языке без изменения (КПСС, СССР, вуз, роно и т. д.), поэтому можно использовать словари аббревиатур и сокращений русского языка (например: [5]), хотя, конечно, останется ряд сокращений, которых не будет в этих словарях, поскольку появились в последние два десятилетия и являются «калмыцкими» по своему происхождению. Например, ***ХТ – Хальмг Таңһч***.

2. Сокращения. Здесь более проблематично, нежели с предыдущим пунктом. Правила сокращений лексических единиц практически не выработаны, это порождает множество вариантов акронимов. Возьмем в качестве примера слово ***җилмүд***: в одних источниках оно сокращается как ***җүж.***, в других как ***ж.ж***., в третьих как ***ж.-ж.*** Следовательно, нужно отыскать в текстах такого рода сокращения и привести их к единообразию. Видимо, это будет осуществляться только уже после работы морфологического анализатора, когда будет известен список неразобранных единиц. Конечно, необходимо выработать правила сокращения слов в тексте.

3. Примыкающие частицы. Целью является отделить слово от примыкающих частиц и корректно тегировать каждый элемент отдельно. Чтобы осуществить это, модуль графематического анализа должен использовать следующую информацию:

a. список частиц и их алломорфов, которые примыкают к слову и становятся, по сути, аффиксами. Здесь же даются и те части речи, к которым может присоединяться определенная частица;

b. при разработке моделей порождения словоизменения в калмыцком языке учитываются и частицы, хотя, оговоримся, что сами частицы не являются аффиксами словоизменения, они лишь придают дополнительный смысл либо самому слову, либо всему предложению. Это своего рода всевозможные комбинации сочетаний различных словоизменительных аффиксов и частиц;

4. Редупликация. В калмыцком языке редупликация является одним из широко распространенных словообразовательных средств. Например, ***хая-хая*** ‘изредка, иногда’, ***таш-таш***, дар-дар. Список таких слов в калмыцком языкознании также отсутствует, поэтому при составлении лексикона старались восстановить эти единицы (в калмыцко-русском словаре Б. Д. Муниева подавляющая масса парных слов приводилась со знаком «:» так же, как и слова, которые в контексте «работают» несамостоятельно [6]).

5. Устойчивые выражения. Здесь в качестве иллюстраций можно привести фразеологическое единство ***улан махмудтан курх*** ‘обнищать’, где три слова и только последнее ***курх*** изменяется, хотя в целом это сочетание слов является одним целым. Решением этой проблемы является создание словаря устойчивых выражений.

6. Композиты и «эквиваленты слов»:

a) самостоятельные части речи, которые в калмыцком языке, как правило, пишутся отдельно и в некоторых случаях через дефис: ***ə-чимən уга*** ‘затишье’, ***мек-тах*** ‘обман’, ***эв-дов*** ‘способ’, ***шикр-микр*** ‘пряности’, ***шар шовун*** ‘филин’, ***темən чикн*** ‘щавель’, ***нег дəкж*** ‘однажды’, ***на-ца уга*** ‘без обиняков’ и др. Каждый из этих элементов обозначает одно понятие, ни при этом структура их компонентна.

b) составные союзы ***тер төлəд***, ***тиигсн бийинь***, ***тедү дүңгə, тедү чигн, гисн кевтə***;

c) сдвоенные союзы ***хəрнь тегəд, болв зуг, декəд бас, бас нам***;

d) повторяющиеся союзы ***аль … аль; эс гиж … эс гиж; негт … негт***;

e) соотносимые союзы ***кен … тер; кедү … тедү; альдаран … тиигəн; хама … тенд; ямаран кевəр … тиим метəр***.

На основе имеющегося опыта других информационно-справочных систем (см.: [7]) мы попытаемся создать список эквивалентов слов. В качестве образца будут взяты словари многокомпонентных единиц как русского и английского, так и турецкого и монгольского языков. Список, приведенный здесь, конечно, неполный, в данный момент грамматический словарь пополняется и новыми элементами, в том числе и композитами.

Таким образом, в работе было кратко описаны проблемы графематического анализа и их возможные решения. Простота, на первый взгляд, графематического анализа калмыцкого текста вместе с тем не отменяет его важности, поскольку является базовым в череде автоматических обработок текстов. Наше решение этих проблем соответствует традициям и схемам, выработанным в корпусной лингвистике.

**Литература**

1. См., например: Муталов Р. О. Корпусная лингвистика и перспективы ее развития в Дагестане // Махачкала: Современные проблемы кавказского языкознания, 2007. Вып. 7. С. 160–173; Он же. Опыт создания корпусов дагестанских языков [Электронный ресурс] // URL: http://www.dialog-21.ru/dialog2009/materials/html/50.htm (дата обращения: 23.04.2011); Бадмаева Л. Д. Бурятский язык и корпусная лингвистика // Состояние и перспективы развития бурятского языка. Материалы форума бурятского языка. – Улан-Удэ, 2009. С. 83–86; Ринчинов О. С. Корпус бурятского языка и прикладные задачи компьютерной лингвистики // Состояние и перспективы развития бурятского языка. Материалы форума бурятского языка. – Улан-Удэ, 2009. С. 88–89; Бадмаева Л. Д. Корпус бурятского языка: проект [Электронный ресурс] // URL: http://www.globecsi.ru/ Articles/2008/Badmaeva3.pdf (дата обращения 20.04.2011) и др.

2. Куканова В. В. Архитектура метаописания в Национальном корпусе калмыцкого языка // Вестник Калмыцкого института гуманитарных исследований РАН. 2011. № 2. С. 139–145.

3. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е. И., Клышинский Э. С., Ландэ Д. В., Носков А. А., Пескова О. В., Ягунова Е. В. – М.: МИЭМ, 2011.

4. Сокирко А. Графематический анализатор [Электронный ресурс] // URL: http://aot.ru/docs/graphan.html (дата обращения 25.11.2012).

5. Новый словарь сокращений русского языка. – М.: ЭТС, 1995 (полностью включает словарь 1984 года). Около 32000 сокращений.

6. Калмыцко-русский словарь / под ред. Б. Д. Муниева. М.: Изд-во «Русский язык», 1977.

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

***

***А.Г. Кукеев,***
*КИГИ РАН*

## КАЛМЫЦКИЕ БОЕВЫЕ ЗНАМЕНА И СЮЖЕТЫ, ИЗОБРАЖЕННЫЕ НА НИХ

Калмыцкие знамёна – интересная тема, выражающая этническое своеобразие традиционной культуры и в определенной степени связанная с традициями иконографии буддизма. Знамя является атрибутом войска, имеющим особое, сакральное значение. Согласно древним представлениям, в знамени живет дух божества-покровителя воинства. Потеря знамени или нанесение ему вреда воспринимались сродни поражению на поле битвы. В калмыцком героическом эпосе «Джангар», богатыри стремятся разорвать знамя противника, либо воткнуть его в землю навершием. Такой поступок, как правило, наводит ужас на врага и лишает его способности к сопротивлению.

В книге Л.А. Боброва и Ю.С. Худякова «Вооружение и тактика кочевников Центральной Азии и Южной Сибири в эпоху позднего средневековья и раннего нового времени (XV-первая половина XVIII в.)»