УДК 81’33
ББК 81.23

НАЦИОНАЛЬНЫЙ КОРПУС КАЛМЫЦКОГО ЯЗЫКА: АРХИТЕКТУРА И ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ*

В. В. Куканова, Е. В. Бембеев, Н. М. Мулаева, Н. Ч. Очирова

Национальный корпус калмыцкого языка является одним из проектов, который позволит сохранить язык в условиях постоянного сокращения числа его активных носителей[1]. Этот программный продукт создаст широкие возможности для проведения лингвистических исследований с использованием корпусного подхода, что не только уточнит и дополнит имеющиеся описания языка, но и осветит те проблемы, которые еще не подняты в калмыцком языкознании.

По мнению Е. И. Кузьмина, председателя Российского комитета Программы ЮНЕСКО «Информация для всех», проблемы сохранения языкового и культурного разнообразия «…приобрели особую актуальность в связи с бурным развитием процессов формирования глобального информационного общества, которые, с одной стороны, усиливают унификацию культур и ведут к сокращению культурного разнообразия, но, с другой стороны, открывают благоприятные возможности для его сохранения и даже для его развития в киберпространстве» [Кузьмин 2010: 40].

В данной статье мы попытаемся осветить актуальность каждого из подкорпусов, его перспективность, проблемы и структуру. По предварительным соображениям необходимо развивать этот ресурс по нескольким направлениям (подкорпусы расположены по приоритетности разработки[2]):

1) основной корпус;
2) газетный подкорпус;
3) устный подкорпус;
4) обучающий подкорпус;
5) параллельный подкорпус;
6) диалектный подкорпус;
7) фольклорный подкорпус;
8) корпус ранних текстов;
9) морфемный подкорпус;
10) поэтический подкорпус;
11) синтаксический подкорпус;
12) подкорпус названий[3].

Создание каждого подкорпуса требует реализации множества предварительных шагов: компилирования словарных материалов, подготовки текстов, разработки концепции каждого подкорпуса, аннотации и разметки.

Источником получения текстов для корпуса калмыцкого языка являются сканированные копии книг, PDF-файлы, файлы верстки (QXD, INDD и др.). Следует признать, что материалов на калмыцком языке, размещенных в сети Интернет, ничтожно мало. Все привлекаемые файлы были преобразованы в формат RTF, необходимый для конвертации в систему StarLing[4]. Были выработаны следующие правила подготовки текстов для корпуса, согласно которым следует:

[1] Под активными носителями понимаются те, кто не только понимает речь, но и осуществляет коммуникацию на том или ином языке, соответственно под пассивными — те, кто только понимает речь окружающих, но, однако, не может продуцировать свои собственные речевые отрезки.

[2] Классифицирование подкорпусов по единому критерию невозможно, поскольку существует дисбаланс некоторых стилей, типов текстов и т. д.

[3] Этот модуль создается на основе метаописания текстов, поэтому это не самостоятельный раздел, а дополнительное поле для исследований (подробно см. ниже).

[4] Информационная среда StarLing создана С. А. Старостиным (1953–2005), а позже усовершенствована Ф. С. Крыловым.

* 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: создание и разработка» (12-04-12047/в).

1) удалять текст, не являющийся авторским (номера страниц, колонтитулы, титульная страница, оборот титула, выходные данные, содержание, аннотация);
2) удалять нетекстовый материал (изображения, схемы, формулы и т. д.);
3) удалять переводные комментарии (как авторские, так и редакторские);
4) упрощать форматирование (шрифтовое выделение заголовка);
5) удалять так называемый «мусор» (двойные пробелы, двойные абзацы, табуляции и т. д.)[5];
6) исправление явных опечаток (две точки вместо одной или две точки вместо трех и т. д.).

Первый ***подкорпус***, по традиции называемый ***основной***, включает художественные, научные и официально-деловые тексты XX в. По статистике базы данных MetaKT[6], прозаические произведения в процентном соотношении составляют 15 % от всего количества произведений, получивших метаразметку. В калмыцкой художественной литературе преобладают тексты преимущественно поэтического характера, которые являются небольшими по своему объему, этот факт обуславливает большой крен в их сторону в плане балансировки текстового материала в создаваемом корпусе. Как известно, в поэтических произведениях язык носит метафорический характер, и слово здесь как таковое переосмыслено автором, что в большинстве случаев создает трудности для описания его лексического значения. Поэтические произведения идеально подходят для изучения идиолекта того или иного автора, а также ритмико-мелодического устройства языка.

Что касается текстов других стилей (научный и официально-деловой), то их не так много, и они составляют небольшую часть имеющихся текстов (менее 1 % от всего массива). Терминология научной сферы разработана достаточно подробно, о чем свидетельствуют терминологические словари калмыцкого языка [Краткий словарь… 1968; Очир-Гаряев 1990; 1995; 1996; Корсункиев 1992; Бардаев 2007; Манджикова 2007]. Однако эти термины не находят должного применения на практике, поскольку вся коммуникация в данных сферах осуществляется на русском языке.

В основном подкорпусе тексты размечаются с морфологической и семантической точек зрения. Ниже приведена система помет для морфологического аннотирования с опорой на разработки: [Овсянникова 2009: 866–871].

Часть речи
Noun — имя существительное
Adj — имя прилагательное
V — глагол
Ptcpl — причастие
Conv — деепричастие
Num — числительное
Adv — наречие
Pron — местоимение
Post — послелог
Conj — союз
Part — частица
Intj — междометие

Число

Sg	единственное число
PL	множественное число
Sgtm	singularia tantum
Pltm	pluralia tantum

Падеж

Nom	именительный падеж
Gen	родительный падеж
Acc1	винительный маркированный падеж
Acc2	винительный немаркированный падеж
Dat	дательно-местный падеж
Inst	орудный падеж
Com	соединительный падеж
Assoc	совместный падеж
Abl	исходный падеж
Dir	направительный падеж
Term	предельный падеж

Прилагательное

Qual.Adj	качественное прилагательное
Rel.Adj	относительное прилагательное
Nmlz	субстантивированный атрибут
Adjz	адъективированное существительное

Числительное

Ord.Num	порядковое числительное
Card.Num	количественное числительное
Par.Num	разделительное числительное
Age.Num	возрастные числительные
Col.Num	собирательное числительное

[5] «Мусор» удаляется автоматически.

[6] База данных MetaKT сконструирована в MS Office Access 2007, в которой дана метаразметка текстов (см. подробно [Куканова 2011]).

Группы местоимений

Pers.Pron	личные местоимения
Dem.Pron	указательные местоимения
Qua.Pron	определительные местоимения
Refl.Pron	возвратные местоимения
Inter.Pron	вопросительные местоимения
Ind.Pron	неопределенные местоимения

Деепричастие

Cond.Conv	условное деепричастие
Term.Conv	предельное деепричастие
Succ.Conv	последовательное деепричастие
Prel.Conv	предварительное деепричастие
Purp.Conv	целевое деепричастие
Conc.Conv	уступительное деепричастие
Prog.Conv	продолжительное деепричастие
Ipfv.Conv	соединительное деепричастие
Ant.Conv	разделительное деепричастие
Mod.Conv	слитное деепричастие

Причастие

Pass.Ptcpl	страдательное причастие
Pos.Ptcpl	причастие возможности
Pres.Ptcpl	причастие настоящего времени
Mom.Ptcpl	однократное причастие настоящего времени
Hab.Ptcpl	многократное причастие настоящего времени
Pst.Ptcpl	причастие прошедшего времени
Fut.Ptcpl	причастие будущего времени

Частица

Refl.Part	рефлексивная частица
Q.Part	вопросительная частица
Pred.Part	предикативная частица (аффикс сказуемости)
Neg.Part	частица отрицания
Mod.Part	модальная частица
Conc.Part	уступительная частица
Conf.Part	подтвердительная частица
Emp.Part	усилительная частица
Equ.Part	уподобительная частица

Категория времени

Pres	настоящее время
Progr	настоящее актуальное время
Past	прошедшее время
Evd	прошедшее результативное
Rem	давнопрошедшее время
Pperf	преждепрошедшее время
Fut	будущее время

Категория лица

1	первое лицо
2	второе лицо
3	третье лицо

Категория наклонения

Indc	изъявительное наклонение
Impr	повелительное наклонение
Hort	желательное наклонение
Juss1	желательное наклонение
Juss2	желательное наклонение
Opt	желательнее наклонение
Appr	предостерегательное наклонение

Категория залога

Pass	страдательный залог
Caus1	побудительный залог
Caus2	побудительный залог
Soc	совместный залог
Recp	взаимный залог

Способы глагольного действия

Dur	длительный вид
Iter	ритмичный вид
Compl	законченный вид
Punc	кратковременный вид
Distr	множественность глагола

Детерминация

Poss	определенность и принадлежность третьему лицу
Poss1	принадлежность первому лицу
Poss2	принадлежность второму лицу

Отрицание

Neg	отрицание

Прочие пометы

Abbr	аббревиатура
ProperN	собственные имена
Name	имя
Surn	фамилия
Patr	отчество
Geox	топоним
Orgn	организация
Non-Standart	нестандартные формы
Dial	диалектная форма
Anom	аномальная форма
Distort	искаженная форма
Ciph	цифровая запись
Init	инициалы
Ab	сокращение

При создании основного подкорпуса использовались материалы издательского дома «Герел», который предоставил электронные копии сверстанных книг. Этот хоть

и небольшой по объему материал стал первым заделом в создании корпуса калмыцкого языка. Однако в электронных копиях мы столкнулись с проблемами кодировки как исконно калмыцких букв, так и всего массива кириллических букв. Для устранения этой проблемы была создана программа Replacer[7], основной целью которой является приведение кодировок калмыцких букв к стандартам UNICODE. Кроме того, программа используется для транслитерации с латиницы на кириллицу и наоборот (например, при разработке Словарного модуля в корпусе).

Другой проблемой, с которой мы столкнулись, является так называемая орфографическая унификация (характерна для всех типов текстов на калмыцком языке, созданных с 1924 г.). Как известно, за этот период сменилось несколько графических систем и орфографических правил, что привело к вариативности написания того или иного слова. Было принято решение сохранять оригинальную орфографию и пунктуацию текста, поскольку этот материал в перспективе также может стать объектом для исследования, ведь становление и развитие норм в калмыцком языке, функционирование узуса еще не были предметом специального изучения. Такое решение повлекло за собой создание словаря вариативных написаний.

Газетный подкорпус. Издательство (редакция) национальной газеты «Хальмг үнн» («*Калмыцкая правда*») любезно передало нам архив за 10 лет (2002–2012 гг.), материал публицистического характера больших объемов и разных жанров. По этой причине было принято решение создать отдельный модуль на основе газетных текстов.

Общественно-политическая газета «Хальмг үнн» основана в 1920 г. под названием «Улан Хальмг», в 1926 г. переименована в «Таңһчин зәңг», а затем название неоднократно менялось. В связи с незаконной депортацией калмыцкого народа (1943–1956 гг.) газета не издавалась, печатные издания возобновили свою работу лишь в 1957 г. после восстановления республики. В развитии публицистического стиля можно выделить несколько этапов: 1) этап становления (додепортационный) — с 1920 по 1943 гг.; 2) постдепортационный этап — с 1957 г. до 1990-х гг., в котором можно выделить два периода (конец 1950–1970-х гг. и 1980-е гг.); 3) переходный период с 1990-х по 2000-е гг.

Наибольший интерес для газетного подкорпуса калмыцкого языка представляют публикации, появившиеся после 1957 г., т. е. после возвращения калмыков из Сибири, в этот период подавляющая часть населения свободно владела калмыцким языком. С 1957 г. до начала 1990-х гг. материалы газеты полностью печатались на калмыцком языке, в ней публиковалось очень много переводных статей, был распространен жанр очерка. Кроме этого, в газете печатались тексты литературных жанров, среди которых нас интересуют только неопубликованные материалы. Надо признать, что в публикациях советского периода преобладает много советизмов, которые ныне перешли в разряд устаревших слов как в русском, так и в калмыцком языках. Но начиная с 1990-х гг. в национальной газете появилась тенденция уменьшения количества текстов на калмыцком языке. Характерной чертой последнего периода стало и чрезмерное употребление интернационализмов, появившихся в калмыцком языке через посредство русского. Другой чертой текущего периода стало издание текстов на религиозную тематику, что ранее не практиковалось в связи с особенностями конфессиональной политики советского государства.

В жанровую классификацию газетных текстов на калмыцком языке входят статья, очерк, интервью и др. По предварительным расчетам объем газетного подкорпуса, в который войдут материалы начиная с 1957 г., составит более 15 млн словоупотреблений. На первоначальном этапе обрабатываются электронные копии газет, исправляется кодировка букв калмыцкого языка. Заметим, что многие шрифты сейчас уже утеряны и быстро восстановить первоначальный вид текста достаточно сложно и трудоемко, поскольку некоторые буквы имеют кодировку, соответствующую другому символу. Например, за кодом пробела закреплена одна из калмыцких букв, а поскольку шрифта уже нет, получается, что тексты с подобной проблемой требуют дополнительных усилий по его выверке с оригиналом.

Все этапы реформирования калмыцкой орфографии отразились на языке публикаций газеты «Хальмг үнн», что вносит дополнительные трудности при их оцифровке. В связи с этим нами разрабатывается словарь вариативных написаний слов калмыцкого языка.

Газетные статьи, с одной стороны, представляют собой чисто лингвистический

[7] Программа создана А. Ю. Каджиевым.

объект исследования, но, с другой стороны, оцифрованные копии газетных статей — это страницы истории, которые отражают общественно-политическую жизнь республики.

Устный подкорпус. Одной из главных целей создания устного подкорпуса калмыцкого языка является документация *живого* калмыцкого языка во всем его многообразии. Поскольку язык находится на грани исчезновения, то, конечно, первым делом необходимо зафиксировать его в том виде, в каком он звучит — это большая кропотливая работа по записи новых текстов и по оцифровке имеющегося материала. Например, в архивах республиканского телевидения[7] и радио[8] хранится большое количество звучащего материала, представленного в основном в аналоговой записи. Только в последнее десятилетие радио и телевидение перешло на цифровое вещание, и соответственно архив стал также носить цифровой характер. Для успешного осуществления данного проекта требуется привлечение материалов республиканского телевидения, которые предстоит конвертировать из аналогового в цифровой. Звуковой файл должен иметь формат WAVE, поскольку он не нарушает частоту звукового материала и сохраняет его характеристики. Подобные звуковые файлы можно использовать даже для изучения фонетических особенностей речи, не говоря о том, что он может выступать в качестве материала для исследования остальных уровней системы.

Создание устного подкорпуса — одна из главных задач лингвистов, поскольку калмыцкий язык как средство познания и коммуникации находится на грани исчезновения и объема записанного материала принципиально недостаточно, а также он не пригоден для изучения языка на фонетическом уровне. Если структуру языка можно восстановить из письменных текстов, то произношение, артикуляцию и так называемые «живые» процессы устной речи невозможно реконструировать по письменным текстам.

Имеющийся записанный материал можно разделить на две группы:

публичная речь, в которой можно выделить подготовленную и неподготовленную:	**непубличная** речь, речь в том виде, в которой она реально функционирует в обществе:
беседа, встреча с читателями, встреча со слушателями, дискуссия, доклад, интервью, комментарий (напр. спортивный), конференция, круглый стол, лекция, пересказ, пресс-конференция, рассказ, репортаж, речь	монолог; беседа; микродиалог: в библиотеке, домашний; разговор: деловой, воспоминание, телефонный; пересказ: разговора, телепередачи; рассказ; спор и пр.

Если первый блок (публичной речи) в той или иной мере существует (необработанный, нерасшифрованный, но он есть), то второй блок отсутствует: нет записей, предназначенных для лингвистического исследования с соблюдением требования разнообразия коммуникативных ситуаций и коммуникативных сценариев[9], балансировки материала по социальным признакам информантов. С учетом того, что с каждым днем количество говорящих уменьшается, то, представляется необходимым направить все усилия на создание второго блока, при этом проблема балансировки материала по социальным критериям уже существует, и тем или иным способом решить ее невозможно: процентное соотношение групп по возрастному признаку отражает реальную языковую ситуацию, сложившуюся в Республике Калмыкия. Записывать речь старшего поколения калмыков, которое в совершенстве владеет калмыцким языком, следует и

[7] В 1967 г. было завершено строительство телецентра в Элисте, первый пробный эфир состоялся 5 ноября 1967 г., а регулярное телевещание началось с 7 ноября [История ТВ Калмыкии].

[8] Радио Калмыкии начало свое вещание в 1935 г. На этом этапе Калмыцкое радио активно сотрудничало с поэтами Х. Сян-Белгиным, С. Каляевым и др., которые выступали с чтением своих стихов, звучали песни в исполнении артистов. Сегодня в архивных фондах радио хранятся многочисленные записи тех лет [История радио Калмыкии].

[9] Под коммуникативным сценарием понимается модель, описывающая нормальную последовательность событий в частном контексте [Shank, Abelson 1977: 248].

по другой причине: скоро их сменит поколение, не владеющее языком в той степени, в которой знает его предыдущее поколение.

При каталогизации записей устной речи требуется учитывать и тематический аспект, характеристику коммуникантов, если таковые имеются, описание звукового файла с указанием качества записи и возможности быстрой расшифровки, а сама структура метаописания должна повторять основные принципы метаразметки художественных текстов: социальные характеристики информанта. Тематический критерий нужен для быстрого поиска необходимых материалов и его отбора и др.

Устный подкорпус калмыцкого языка можно будет использовать в преподавании разных лингвистических дисциплин как для школьников, так и для студентов-филологов. Данный подкорпус — это реализация реконструкции живой калмыцкой речи, демонстрирующей разные речевые процессы.

Обучающий подкорпус. С учетом опыта функционирования обучающего корпуса русского языка [Добрушина 2005; 2009; Савчук, Сичинава 2009; Образовательный портал Национального корпуса русского языка] предлагается создание обучающего подкорпуса калмыцкого языка со снятой омонимией, разметка которого будет ориентирована на школьную программу калмыцкого языка, т. е. грамматическая информация будет соответствовать, современной школьной программе (например, будут использоваться традиционные лингвистические термины). Предполагается, что, помимо текстов со стандартной грамматической разметкой, обучающий подкорпус калмыцкого языка представит возможность создания упражнений по темам школьной программы и для организации самостоятельной работы школьников и студентов (склонение существительных, спряжение глаголов, образование причастий и деепричастий и др.).

В обучающий подкорпус войдут только те произведения, которые проходят в школе, предполагается, что часть литературных и фольклорных произведений будет доступна для чтения в свободном доступе. Другим критерием для отбора произведений будет выступать степень их кодифицированности, поскольку в школе преподавание ориентировано на литературный язык. В меньшей степени будут представлены диалектные тексты, которые познакомят тех, кто изучает калмыцкий язык, с диалектными особенностями языка. К тому же тексты следует подбирать по разным тематикам для того, чтобы было представлено все лексическое многообразие калмыцкого языка, а это, как известно, самое сложное в обучении иностранному языку, каковым сейчас является калмыцкий язык для большинства детей. Следовательно, в обучающем подкорпусе должны быть представлены упражнения на базе предложений не только грамматического, но и лексического характеров, направленных на закрепление материала, пройденного самостоятельно или с учителем.

Первоначально в обучающий подкорпус калмыцкого языка будут включены несложные упражнения, например: найти в тексте причастие, определить главные и второстепенные члены и т. д. В подкорпус войдут инструкции для учителей по составлению упражнений, а также будут проведены специальные обучающие семинары.

Параллельный подкорпус. Этот подкорпус имеет большое значение для возрождения калмыцкого языка в обществе: большинство детей в основном сначала овладевают русским языком, который становится первичным языком. В процессе изучения языка школьники и студенты опираются на знания первичного языка и изучают калмыцкий язык сквозь призму русского языка, который относится к другому типу — флективному, со свободным порядком слов в предложении. Важно создать ресурс, который поможет сопоставить два разноструктурных языка и вывести соответствия на лексико-грамматическом уровне.

Структура параллельного подкорпуса как самостоятельного модуля в корпусе предварительно будет следующей: русский → калмыцкий, калмыцкий → русский, калмыцкий → английский, монгольский → калмыцкий, калмыцкий → монгольский. Текстов, переведенных с бурятского или на бурятский, пока еще не было обнаружено.

Все тексты требуется отсканировать и распознать, что несколько задерживает работу. В случае с первым модулем русский → калмыцкий гораздо легче, так как требуется распознать только калмыцкие тексты, тексты же на русском языке доступны из Интернета. В остальных модулях сканируются и распознаются книги и с той, и с другой стороны (переводные тексты и тексты-оригиналы). В корпусе переводов с монгольского на калмыцкий пока не найдены их оригина-

лы. Электронных копий переводных текстов крайне мало. Некоторые из них были представлены Р. М. Ханиновой, за что мы выражаем ей огромную благодарность.

В связи с описанными трудностями работа по созданию параллельного подкорпуса была начата с первого модуля. Отсканирована и распознана большая часть переводных произведений: это в основном классическая литература, а также литература социалистического периода. Все эти тексты прошли первичную обработку.

Первым текстом, который был выровнен по предложениям, стала повесть А. С. Пушкина «Капитанская дочка», она была выбрана по нескольким причинам как субъективного, так и объективного характеров: во-первых, произведение небольшое по своему объему (чуть менее 29 500 словоупотреблений), во-вторых, язык повести относится к классическому периоду, в-третьих, быт периода, который описывается в тексте А. С. Пушкина, не характерен для калмыцкого общества того времени (многих лексических единиц до сих пор нет в системе языка), в-четвертых, имя А. С. Пушкина пользуется особой любовью калмыцкого народа.

Диалектный подкорпус. Диалектная система калмыцкого языка представляет собой совокупность трех территориально-языковых разновидностей: дербетского, бузавского и торгутского говоров [Кичиков 1963; Бардаев 1985; Убушаев 2006 и др.]. Создание диалектного подкорпуса в Национальном корпусе калмыцкого языка даст возможность сравнивать диалекты с литературным калмыцким языком, в частности, позволит выяснить соотношение частотности диалектных явлений, а также изучать их грамматические свойства и т. д.

Проанализировав исследования по созданию и развитию диалектных корпусов [Летучий 2005; 2009; Крючкова, Гольдин 2008; 2011; Некрасова 2009; Юрина 2011], можно выделить два вида диалектных корпусов: 1) информационно-справочная система, где целью аннотирования является выделение ненормированных языковых фактов в текстах; 2) информационно-справочная система, содержащая диалектные тексты в фонетической транскрипции. Первый тип не интересен диалектологам, поскольку этот материал, в котором тот или иной лингвист уже констатировал наличие диалектных явлений. Второй, напротив, является материалом, который еще не изучен ни в каких-либо аспектах и который может быть источником для уточнения сведений по уже известным фактам и для изучения новых диалектных явлений, характерных для определенной территориальной разновидности языка.

Было решено, что в основном подкорпусе будут помечаться особенности диалектной системы, поскольку на данный момент в нашем распоряжении находятся только письменные тексты, в которых можно найти большое количество примеров отступлений от нормы: важно обозначить эти некодифицированные элементы калмыцкого языка, чтобы пользователь видел проявление нормы и узуса. Конечно, тем самым сузится поле для исследователей, прежде всего для диалектологов. Если в диалектном подкорпусе Национального корпуса русского языка выбирается морфологически ориентированная стратегия: «отмечаются только те отличия от литературного языка, которые имеют отношения к грамматике или отражаются на грамматических особенностях» [Летучий 2005: 217], то в калмыцком языке невозможно избрать только данную стратегию, поскольку большинство диалектных различий приходятся на фонетику и лексику. Морфологических особенностей не так много, хотя они имеются. Например, при выделении словоизменительных классов помечались диалектные варианты словоизменений (см. подробнее [Куканова 2012а]). Так, плюральный аффикс *-дуд/дүд*, который является составным (*-д* + *-уд/үд*), встречается в торгутском диалекте и образует множественное число у существительных на неустойчивый *-н*, когда в литературном языке у таких слов исчезает -н и присоединяется аффикс *-д* [Убушаев 2006: 7]. Были выделены три типа диалектных особенностей в калмыцком языке:

1) *dialfon* — особенности на фонетическом уровне, отражающиеся в письменной форме;
2) *diallex* — особенности на лексическом уровне;
3) *dialmorf* — особенности на морфологическом уровне.

Для каждого говора введен индекс: T — торгутский, D — дербетский, B — бузавский. Пометы *dialfon* и *dialmorf* конкретизируются: указывается, в чем состоит диалектная особенность (подробнее см. примеры разметки в: [Куканова, Очирова 2012]).

В перспективе отдельно будет создаваться диалектный подкорпус, основан-

ный на фонетических расшифровках. Расшифровка будет представлена по правилам Международного фонетического фонда, что, следовательно, сделает доступными диалектные тексты как для отечественных, так и зарубежных исследователей. В архивах Калмыцкого института гуманитарных исследований РАН содержится большое количество аудиозаписей, однако они не расшифрованы в нужном для диалектологов виде. Требуется разработка инструкции по расшифровке звукового материала и программа записи текстов. Видимо, следует записывать тексты на заданные темы, чтение текстов не является достоверным материалом в этом случае, поскольку письменный текст оказывает большое влияние на читающего. В текстах-монологах на заданную тему можно получить относительно чистый материал для изучения диалектных черт, поскольку говорящий чувствует определенную свободу в порождении текста, а одна и та же тема позволит сравнивать одинаковые элементы в речи носителей.

Фольклорный подкорпус. Данный подкорпус будет состоять из прецедентных текстов — фольклорных произведений, являющихся неотъемлемой составляющей духовного наследия калмыцкого народа. В них отражены древнее мировоззрение и мироощущение народа, наивная картина мира во всех своих категориях, универсалиях и специфических чертах, например понятия времени и локации, персональности и движения и многое другое. Фольклор во всем своем многообразии жанров являет собой яркий образец метафоричности языка и содержит элементы архаики, по этой причине (и не только), представляется важным включение фольклорных произведений в Национальный корпус калмыцкого языка.

Создана структура базы данных, в которой отражено метаописание фольклорных текстов [Куканова 2012б]. Эта база данных носит не собственно лингвистический характер, а направлена прежде всего на сохранение фольклорного наследия калмыцкого этноса, поэтому специально рассматривать этот подкорпус в данной статье мы не будем.

Корпус «ранних» текстов. Этот модуль является одним из самых трудновыполнимых по нескольким причинам: в первую очередь отсутствуют тексты на «тодо бичиг», из которых можно было бы получить объемный словарь грамматических форм; отсутствует юникодовая кодировка всех графем, пишущихся в середине и в конце слова; отсутствует поддержка вертикального письма в текстовых редакторах; отсутствует распознающая программа, которая бы облегчила подготовку текстов на «тодо бичиг» [см. подробно Бембеев 2012а; 2012б].

Тем не менее было принято решение транслитерировать тексты на латиницу, кроме этого, ведется работа по созданию грамматического словаря старокалмыцкого языка: сейчас сформирован словник, состоящий из материалов двух словарей [см. подробно Мулаева 2012; Очирова 2012]. Работа по созданию корпуса «ранних» текстов необходима для реконструкции старокалмыцкого языка: этот период не исследован системно и глубоко, изучены лишь отдельные факты и жанры текстов (в частности [Сусеева 2003; Гедеева 2004]).

Д. А. Павлов выделяет три этапа в становлении современного калмыцкого языка. Второй этап делится на два: конец XIV и до первой половины XVII вв. и вторая половина XVII в. до 1917 г. [Павлов 2000]. В архивах содержатся памятники XVII в. и начала XX в. Период достаточно большой, что привело нас к разграничению его на составляющие: материал собирается по каждому веку отдельно. По жанровой представленности говорить пока еще рано, но очевидно, что это письма, историко-литературные памятники, религиозные произведения разных жанров (притчи, хождения, трактаты и т. д.).

Синтаксический подкорпус. Существует два вида синтаксического корпуса: 1) корпус, разметка в котором основана на выделении синтаксических ролей и характеристике словосочетания, клаузы, предложения с различных точек зрения (см., например, корпус ХАНКО [Копотев, Мустойоки 2003]); 2) синтаксически и семантически аннотированный корпус (основан на теории Смысл ↔ Текст, разработанной И. А. Мельчуком [1999]). В последнем представлено дерево зависимостей [Апресян и др. 2005].

На наш взгляд, было бы интересно приложить к нашему проекту теорию Смысл ↔ Текст и разработать синтаксически и семантически аннотированный корпус. Корпус будет, скорее всего, небольшим по своему объему, так как это один из самых сложных и трудновыполнимых проектов.

Морфемный подкорпус. Это один из наиболее интересных проектов для изучения морфемного состава калмыцкого слова.

Данный подкорпус позволит исследовать структуру агглютинативного слова на основе корпусного подхода и создать описание значений морфем, начиная с частотных и заканчивая нечастотными элементами. Поскольку калмыцкий язык относится к монгольской группе языков, то гипотетически будет несложно технически реализовать эту задачу. В связи с этим необходимо будет создать словарь морфем и словарь слов с морфемным членением. В качестве исходного материала будет выступать словник из грамматического словаря, основанный в свою очередь на словнике Калмыцко-русского словаря под ред. Б. Д. Муниева [1977].

Структура первого словаря должна выглядеть следующим образом (см. таблицу 1):
1) ID;
2) морфема;
3) тип морфемы:
- корень;
- аффикс;

4) производящая основа и производная основа:
- аффикс, образующий глагол от глагольной основы;
- аффикс, образующий глагол от основы прилагательного;
- аффикс, образующий глагол от основы существительного;
- аффикс, образующий глагол от основы числительного;
- аффикс, образующий глагол от основы идеофонов;
- аффикс, образующий имя прилагательное от основы существительного;
- аффикс, образующий имя существительное от глагольной основы;
- аффикс, образующий наречия от основы имени существительного, и др.

Структура второго словаря будет состоять из: 1) ID; 2) слово; 3) часть речи; 4) структура слова (см. таблицу 2).

Таблица 1. ***Структура словаря морфемного членения слов калмыцкого языка***

ID	Морфема	Тип морфемы	Производящая основа	Производная основа	Пример
1	*-лһн*	Aff	V	N	*умш-лһн* ‘чтение’
2	*-лһ*	Aff	V	N	*ав-лһ* ‘взятка’
3	*-ч*	Aff	N	N	*үкр-ч* ‘пастух’
4	*-ач*	Aff	V	N	*умш-ач* ‘читатель’
5	*-әч*	Aff	V	N	*бич-әч* ‘писатель’

Таблица 2. ***Структура словаря морфем калмыцкого языка***

ID	WORD	POS	StrWord
22639	*чидл*	N	*чид-л*
24304	*эвцлһн*	N	*эвц-лһн*
3726	*бүсл-*	V	*бүс-л-*

Оговоримся сразу, аффиксы залога также будут рассматриваться в морфемном словаре, поскольку данные форманты обладают и словообразовательной функцией. Помимо этого, нужно также создать свод морфонологических правил, которые действуют на стыке аффиксов, а также корня и аффиксов, регулируя соединение морфемных элементов в слове.

Таким образом, создание и разработка морфемного подкорпуса имеет большое значение, поскольку традиционно считается, что агглютинативные языки рассматриваются как языки с традиционно бедной морфонологией [Грунтов 2006: 148]. 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 [Крылов 2004]. Поскольку халха-монгольский и калмыцкий языки родственны, 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» [Трубецкой 1967: 119].

Поэтический подкорпус. Поэтические тексты, наравне с прозаическими, диалектными, фольклорными, являются весьма важными источниками изучения калмыцкого языка. Поэтический подкорпус калмыцкого языка во многом ориентирован на разработки, предложенные в работе [Гришина и др. 2009]. Создатели Национального корпуса русского языка отмечали важность присутствия в нем представительного электронного массива поэтических текстов. Сложность разметки этих текстов была причиной того, что работа над поэтическим корпусом началась не сразу, а лишь после того, как основной корпус (прозаические художественные и нехудожественные тексты) достиг более ста миллионов словоупотреблений и основные принципы метатекстовой и морфологической разметки стали более ясны [Гришина и др. 2009: 72].

Данный подкорпус весьма важен для изучения ритмико-мелодической системы языка, его потенциала. Тем не менее необходимо сначала исследовать ритмику и мелодику стихотворной калмыцкой речи, по этой причине данный подкорпус по приоритетности стоит на последнем месте. Помимо морфологической и семантической разметки, тексты будут сопровождаться и специальной разметкой, отражающей особенности ритмико-тонической организации стихотворения.

На данном этапе можно выявить только особенности структурной организации строфы и рифмы. В первом случае классификация опирается на классические литературоведческие работы по стихосложению: [Гаспаров 2001]. Рифмовка в калмыцком поэтическом произведении оригинальна, поскольку строки рифмуются не только по концу, но и по началу строки.

В рамках работы над проектом будут размечены поэтические произведения калмыцких авторов по трем периодам: 1) 1920–1940-е гг., ранняя советская поэзия; 2) 1957–1980-е гг. и 3) 1990–2000-е гг., современная поэзия. Обращение к творчеству поэтов разных периодов позволит специалистам, работающим с литературными поэтическими текстами, существенно уточнить ряд особенностей литературы рассматриваемых периодов.

Подкорпус названий. Заголовки — «это тексты второго порядка: метатекст по отношению к основному тексту и одновременно просто текст как таковой и небольшого объема» [Гришина 2005: 246]. Однако под названиями мы понимаем только заголовки текстов, в отличие от Е. А. Гришиной, которая включает следующие объекты:

- названия артефактов: заголовки текстов и ярлыки (названия учреждений, объектов культуры и т. д.);
- названия природных объектов [Гришина 2005: 244].

Всевозможные онимы в корпусе получат свою разметку при семантическом аннотировании, система которой разрабатывается в настоящий момент. Материалом послужит созданная база данных MetaKT, объем которой уже репрезентативен для проведения исследований по заголовкам текстов, написанных на калмыцком языке, при этом она постоянно пополняется новыми материалами.

Метаописание для данного подкорпуса строится на тех же самых принципах и, более того, содержится в той же самой базе данных. Перспективы использования данного проекта достаточно широки, и его можно реализовывать параллельно с другими модулями, поскольку текстовый материал и метаразметка уже готовы.

Таким образом, Национальный корпус калмыцкого языка будет состоять из нескольких модулей, о которых было сказано выше, однако его разработчики не забывают о приоритетности отдельных его модулей, поскольку необходимость создания некоторых диктуется самой языковой ситуацией, актуальностью создания «Академической грамматики калмыцкого языка» на основе корпусного подхода, как правило, учитыва-

ющего все стороны и сферы языка. В силу этого приоритетными являются основной, устный, параллельный и обучающий подкорпусы.

Понятно, что на создание последних четырех потребуется гораздо больше усилий и затрат, чем на первый. Но, с другой стороны, для реализации основного подкорпуса нужно выполнить много «ручной» работы, а именно филологической выверки большого массива текстов, их орфографической унификации (о специфике унификации см. выше), данную работу можно лишь частично автоматизировать. Поэтому совершенно ясно, что в ближайшие годы разработка подкорпусов будет вестись выборочно и, конечно, будет во многом зависеть от потребностей языкового общества. Реализация проекта «Национальный корпус калмыцкого языка» позволит решить множество лингвистических задач[10], которые на начальном этапе создания корпуса еще даже сложно очертить и объективно оценить. В лингвистике первый этап всегда самый сложный, трудоемкий и занимающий много времени, но результаты подготовки материала для корпуса позволят сэкономить время и усилия для будущей исследовательской работы.

Литература

Schank R. C., Abelson R. P. Scripts, Plans, Coals and Understanding: An Inquiry into Human Knowledge Structures. Hillsdale, N.J.: Lawrence Erlbaum Ass., 1977. 248 p.

Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л., Санников А. В., Санников В. З., Сизов В. Г., Цинман Л. Л. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка 2003–2005 гг. (результаты и перспективы). М.: Индрик, 2005. С. 193–214.

Бардаев Э. Ч. Материалы к калмыцко-русскому и русско-калмыцкому словарю лингвистических терминов. Элиста: КИГИ РАН, 2007. 102 с.

Бардаев Э. Ч. Современный калмыцкий язык. Лексикология / под ред. Г. Ц. Пюрбеева. Элиста: Калм. кн. изд-во, 1985. 154 с.

Бембеев Е. В. Коллекции рукописей на старокалмыцком (ойратском) языке XVII–XIX вв. в свете компьютерной обработки: постановка проблемы // Информационные технологии и письменное наследие. El'Manuscript-2012: Мат-лы IV Междунар. науч. конф. (Петрозаводск, 3–8 сентября 2012 г.). Петрозаводск, Ижевск, 2012а. С. 31–34.

Бембеев Е. В. Опыт квантитативной обработки текста на старокалмыцком языке: количественные характеристики // Вестник Калмыцкого института гуманитарных исследований РАН. 2012б. № 2. С. 163–168.

Гаспаров М. Л. Русский стих начала XX века в комментариях. Изд. 2-е (доп.). М.: Фортуна Лимитед, 2001. 288 с.

Гедеева Д. Б. Письма наместника Калмыцкого ханства Убаши (XVIII в.) / отв. ред. Э. У. Омакаева. Факсимиле писем. Изд. текстов, введение, транслитер., пер. со старокалм. на совр. калм. яз., сл. / Д. Б. Гедеева. Элиста: АПП «Джангар», 2004. 196 с.

Гришина Е. А., Корчагин К. М., Плунгян В. А., Сичинава Д. В. Поэтический корпус в рамках НКРЯ: общая структура и перспективы использования // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 71–113.

Гришина Е. А. Два новых проекта для Национального корпуса: мультимедийный подкорпус и подкорпус названий // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 233–250.

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.

Добрушина Н. Р. Как использовать Национальный корпус русского языка в образовании? // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 308–329.

Добрушина Н. Р. Корпусные методики обучения русскому языку // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 335–352.

[10] Надо сказать, что выполнение проекта выходит за пределы только лингвистических задач (см. к примеру фольклорный или обучающий подкорпус, на основе которых в первом случае можно исследовать проблемы фольклористики, а во втором — проблемы методики преподавания калмыцкого языка в условиях его исчезновения).

История радио Калмыкии [электронный ресурс] // URL: http://vesti-kalmykia.ru/tv_history.html копия (дата обращения: 27.08.2012).

История ТВ Калмыкии [электронный ресурс] // URL: http://vesti-kalmykia.ru/tv_history.html (дата обращения: 27.08.2012).

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

Кичиков А. Ш. Дербетский говор: автореф. дис. … канд. филол. наук. М.; Элиста, 1963. 26 с.

Копотев М. В., Мустайоки А. Принципы создания Хельсинского аннотированного корпуса русских текстов (ХАНКО) в сети Интернет // Научно-техническая информация. Сер. 2. Информативные процессы и системы. 2003. № 6. С. 33–37.

Корсункиев Ц. К. Калмыцко-русский и русско-калмыцкий терминологический словарь: Медицина. Элиста: КИГИ РАН, 1992. 190 с.

Краткий словарь общественно-политических терминов калмыцкого языка. Элиста, 1968. 88 с.

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.

Крючкова О. Ю., Гольдин В. Е. Корпус русской диалектной речи: концепция и параметры оценки [электронный ресурс] // URL: http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/36.pdf (дата обращения: 30.08.2012).

Крючкова О. Ю., Гольдин В. Е. Текстовый диалектологический корпус как модель традиционной сельской коммуникации [электронный ресурс] // URL: http://www.dialog-21.ru/digests/dialog2008/materials/html/41.htm (дата обращения: 30.08.2012).

Кузьмин Е. И. Сохранение языкового и культурного разнообразия в России: проблемы и перспективы // Языковое и культурное разнообразие в киберпространстве: мат-лы Междунар. конф. (Якутск, 2–4 июля 2008 г.). М.: МЦБС, 2010. С. 40–51.

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

Куканова В. В. Фольклорный подкорпус: проблемы, структура и перспективы использования // Участие калмыков в укреплении Российской государственности: Мат-лы Регион. науч.-практ. конф., посвящ. 1150-летию Российской государственности и Году российской истории (г. Элиста, 29 ноября 2012 г.). Элиста: КИГИ РАН, 2012. С. 192–197.

Куканова В. В., Очирова Н. Ч. Общее или индивидуальное, норма или узус в Национальном корпусе калмыцкого языка: к постановке проблемы // Актуальные проблемы диалектологии языков народов России: Мат-лы XII Регион. конф. (Уфа, 27–28 ноября 2012 г.). Уфа: УНЦ РАН, 2012. С. 90–94.

Летучий А. Б. Диалектный корпус: состав и особенности разметки // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы / отв. ред. В. А. Плунгян. СПб.: Нестор-История, 2009. С. 114–128.

Летучий А. Б. Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: Индрик, 2005. С. 215–232.

Манджикова Б. Б. Калмыцко-русский терминологический словарь: флора и фауна. Элиста: КИГИ РАН, 2007. 95 с.

Мельчук И. А. Опыт теории лингвистических моделей «Смысл ↔ Текст». М.: Школа «Языки русской культуры», 1999. I–XXII, 346 с. (Сер. Язык, семиотика, культура).

Мулаева Н. М. Русско-калмыцкий словарь Пармена Смирнова как источник изучения лексики калмыцкого языка // Участие калмыков в укреплении Российской государственности: Мат-лы Регион. науч.-практ. конф., посвящ. 1150-летию Российской государственности и Году российской истории (г. Элиста, 29 ноября 2012 г.). Элиста: КИГИ РАН, 2012. С. 187–191.

Некрасова Г. А. Электронный диалектный корпус как ресурс сохранения и изучения коми диалектов // Языковая палитра. 2010. С. 13–16.

Образовательный портал Национального корпуса русского языка [электронный ресурс] // URL: http://studiorum.ruscorpora.ru/index.php?o (дата обращения: 20.08.2012).

Овсянникова М. А. Грамматические показатели калмыцкого языка // Исследования по грамматике калмыцкого языка / Ред. С. С. Сай, В. В. Баранова, Н. В. Сердобольская (ACTA LINGUISTICA PETROPOLITANA. Труды Института лингвистических исследований РАН, 2009. Т. V, ч. 2). СПб.: Наука, 2009. С. 866–872.

Очир-Гаряев В. Э. Калмыцко-русский и русско-калмыцкий словарь терминологический словарь: Агрономия. Элиста: КИГИ РАН, 1990. 85 с.

Очир-Гаряев В. Э. Калмыцко-русский и русско-калмыцкий терминологический словарь: Рыбное хозяйство. Элиста: КИГИ РАН, 1995. 64 с.

Очир-Гаряев В. Э. Калмыцко-русский, русско-калмыцкий терминологический словарь: Народное образование. Элиста: КИГИ РАН, 1996. 91 с.

Очирова Н. Ч. «Ранние» словари калмыцкого языка и современные информационные технологии // Участие калмыков в укреплении Российской государственности: Мат-лы Регион. науч.-практ. конф., посвящ. 1150-летию Российской государственности и Году российской истории (г. Элиста, 29 ноября 2012 г.). Элиста: КИГИ РАН, 2012. С. 183–186.

Павлов Д. А. Формирование и развитие калмыцкого национального литературного языка // Павлов Д. А. Вопросы истории и строя калмыцкого литературного языка. Сб. науч. ст. Изд. 2. Элиста: Калм. гос. ун-т, 2000. С. 17–37.

Савчук С. О., Сичинава Д. В. Обучающий корпус русского языка и его использование в преподавательской практике // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 317–334.

Сусеева Д. А. Письма Аюки-хана и его современников (1714–1724 гг.): опыт лингвосоциологического исследования. Элиста: АПП «Джангар», 2003. 423 с.

Трубецкой Н. С. Некоторые соображения относительно морфонологии // Пражский лингвистический кружок. М.: Прогресс, 1967. С. 115–119.

Убушаев Н. Н. Диалектная система калмыцкого языка / отв. ред. Э. У. Омакаева. Элиста: Джангар, 2006. 256 с.

Юрина Е. А. Томский диалектный корпус: в начале пути // Вестник Томского государственного университета. 2011. № 2(14). С. 58–63.