Электронный информационный журнал
«НОВЫЕ ИССЛЕДОВАНИЯ ТУВЫ»
www.tuva.asia № 1 2015

НАЦИОНАЛЬНЫЙ КОРПУС КАЛМЫЦКОГО ЯЗЫКА: ИТОГИ РАБОТЫ И ПЕРСПЕКТИВЫ

В. В. Куканова

Аннотация: В статье описаны итоги трехлетней работы по проекту «Национальный корпус калмыцкого языка: создание и разработка», поддержанному РГНФ в 2012–2014 г.

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

Ключевые слова: корпусная лингвистика, лингвистические базы данных, компьютерные технологии, архитектура систем обработки естественных языков, метаописание, токенизация, сегментация, лемматизация, морфологическая модель языка, калмыцкий язык.

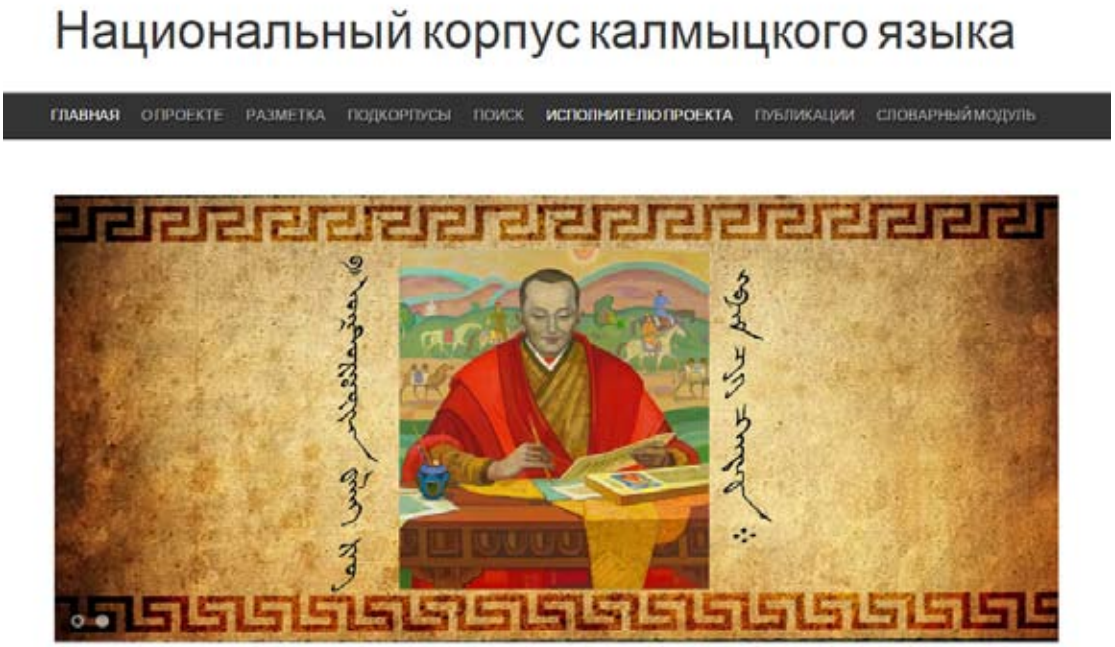
Национальный корпус калмыцкого языка представляется одним из фундаментальных проектов, который сможет внести свою лепту в дело сохранения языка в условиях постоянного уменьшения количества его активных носителей, под которыми понимаются те, кто не только понимает речь, но и осуществляет коммуникацию на том или ином языке. Этот программный продукт позволяет осуществлять лингвистические исследования с использованием корпусного подхода, что не только уточняет и дополняет имеющиеся описания языка, но и освещает те проблемы, которые еще не изучены в калмыцком языкознании.

Основная цель проекта «Национальный корпус калмыцкого языка: создание и разработка» состояла в разработке структуры представительного корпуса калмыцкого языка, его состава, оцифровке текстов. Информационно-справочная система должна выступить как материал для исследования лексики и грамматики калмыцкого языка. Данный про-

Куканова Виктория Васильевна - кандидат филологических наук, заведующая лабораторией прикладной и экспериментальной лингвистики Калмыцкого института гуманитарных исследований РАН, г. Элиста.

ект был фактически начат во второй половине 2012 г., поскольку он получил одобрение Совета РГНФ по дополнительному конкурсу, результаты которого стали известны в июле 2012 г. Этот факт значительно повлиял на сроки реализации задач проекта на 2012 г., поставленных в исходной заявке 2011 г. Проект получал поддержку РГНФ на протяжении трех лет и только благодаря этому был реализован.

Осуществление проекта по созданию Национального корпуса калмыцкого языка предусматривало реализацию нескольких задач: 1) проведение каталогизации существующих текстов и записей на калмыцком языке; 2) оцифровка текстов; 3) обработка электронных файлов на предмет филологической сверки с авторским текстом; 4) метаразметка текста, описание дополнительных параметров текста (метаданные автора и самого текста, т. е. информация экстралингвистического характера); 5) автоматическая обработка текста (токенизация, лемматизация).



В результате его выполнения главным лингвистическим результатом, на наш взгляд, стало создание информационно-справочной системы, основанной на оцифрованном массиве текстов на калмыцком языке, которые сопровождаются лингвистической и металингвистической информацией (**http://kalmcorpora.ru**). Разработанный корпус имеет одну важную особенность: он основан на языке, который находится в условиях постепенной утраты в силу исторических причин (насильственной депортации, потери устойчивой связи поколений в трансляции культурных ценностей, в их чис-

ле и языка как достояния культуры народа). Этот факт еще более усиливает актуальность проведенной работы в этом направлении.

В ходе выполнения проекта был проанализирован опыт отечественной и зарубежной лингвистики по созданию национальных корпусов, объемных текстотек (преимущественно языков агглютинативной структуры). В первую очередь внимание уделялось вопросам репрезентативности текстового материала в свете исчезающих языков. Текстов на калмыцком языке достаточно мало. Объемы их, конечно, не сопоставимы, предположим, с существующим массивом текстов на русском языке, который имеет древнюю письменную традицию. Несмотря на то, что калмыки стали частью Российского государства сравнительно недавно (более 400 лет тому назад) и в этот период у них уже была своя собственная письменность, тем не менее текстов в количественном отношении немного. На графической системе «тодо бичиг» создана обширнейшая литература, но поскольку в конце 1930-х гг. произошла окончательная смена графики на кириллицу, нужно признать, что в этом плане новая графическая система не сыграла роли преемственности передачи письменного наследия последующим поколениям. Для увеличения корпуса было принято решение оцифровать все, что имеется. Конечно, при этом определялась приоритетность источников в сканировании и распознавании.

В результате проведенной работы, говоря об особенностях корпуса, то можно сказать, что созданный электронный ресурс по калмыцкому языку не обладает сбалансированностью, как, например, Национальный корпус русского языка хотя можно утверждать это с известной долей условности.

Причина несбалансированности Национального корпуса калмыцкого языка кроется в существующей языковой ситуации с преобладанием русского языка как средства мышления, познания, общения, не позволяющей создать корпус, который был бы сбалансирован в жанровом, стилистическом и хронологическом планах. Вместе с тем следует отметить, что тексты всех жанров и стилей представлены в корпусе, хронологические рамки оцифрованных текстов достаточно широкие, однако процентное соотношение текстов по разным стилям, жанрам и времени создания не равнозначное. Не все стили представлены равномерно: основной массив текстов — это художественные произведения, затем идут фольклорные тексты, которые относятся к художественному стилю; тексты делового и научного стиля составляют несущественное количество (около 20 текстов на весь массив оцифрованных текстов). В хронологическом плане тексты также не сбалансированы: в корпусе тексты относятся преимущественно

ко второй половине XX в. Представлены тексты начала XVIII в.: это письма ханов, написанные на старокалмыцкой графической системе («тодо бичиг»). Но, поскольку мы отбирали материал методом сплошной выборки, то пока обработано всего 7 первых дел фонда И–36 Национального архива Республики Калмыкия. Тексты, относящиеся ко второй половине XVIII в. и XIX вв. не представлены вовсе. Несбалансированность существует и среди разных форм речи: соотношение прозаических и поэтических художественных текстов составляет 1 к 6. Таким образом, корпус не сбалансирован, но в то же время он отражает реальное функционирование калмыцкого языка в обществе на всем протяжении его истории. Действительно в обществе не сложилась традиция создания деловых документов или научной продукции на калмыцком языке. Существует лишь небольшое количество научных статей, написанных на калмыцком языке: ученые предпочитают писать на русском языке, чтобы быть доступнее более широкому кругу читателей.

Положительной стороной созданного корпуса является прежде всего то, что тексты снабжены разметкой: фонетической, грамматической, семантической и металингвистической. Созданные запросы позволяют лингвистам проводить поиск материала по заданным критериям. В 2014 г. создан полноценный поиск на сайте корпуса калмыцкого языка (ww.kalmcorpora.ru). Если в 2013 г. была запущена простая система поиска по параметрам: «равно», «содержит», «начинается с ...», «заканчивается на ...», то в этом году был реализован расширенный поиск, который имеет уже более сложную структуру: с одной стороны, это поиск по словоформе, лемме или переводу, что традиционно для языковых корпусов, с другой стороны — по грамматическим, лексико-семантическим и дополнительным признакам. Поиск реализован средствами языка запросов базы данных MySQL, что обеспечивает быстродействие их выполнения. Разметка позволяет получить более конкретный материал для проведения исследований. Именно ее наличие отличает корпус от электронной библиотеки текстов. Если на русском языке существует большое количество библиотек, то на калмыцком языке полноценной электронной коллекции текстов нет. Этот фактор создал сложности в создании Национального корпуса калмыцкого языка.

Корпус не возможен, как известно, без текстов, поэтому работа началась с оцифровки и распознавания текстов на калмыцком языке. На наш взгляд, более значительным в культурном плане результатом является работа по проведению оцифровки книг на калмыцком и русском язы-

ках, которые с полным правом можно отнести к письменному наследию калмыцкого этноса. За весь срок выполнения проекта оцифровано более 500 изданий, большая часть из которых на калмыцком языке, а меньшая представляет их переводы. За короткий срок при неоценимой помощи студентов, учителей калмыцкого языка и литературы была проведена кропотливая и сложная работа по проверке текстов после автоматического распознавания текстов, содержащих, несмотря на создание пользовательского словаря для программы ABBYY FINEREADER 11 Pro, множество ошибок как системного, так и несистемного (случайного) характера. Упомянутый пользовательский словарь для распознающей программы и режим обучения позволили уменьшить количество ошибок при автоматическом распознавании, что значительно уменьшило количество ошибок в текстах на калмыцком языке и, следовательно, повысило качество распознавания.

В 2012 г. первоначально тексты сканировались в программе ABBYY FineReader 11 Pro и здесь же проходило исправление ошибок, но впоследствии при выгрузке текстов мы столкнулись с проблемой их конвертации в текстовый формат: появлялись ошибки, связанные с дефисами и мягкими переносами, принудительными разрывами строк, которые отсутствуют в автоматически распознанных копиях текстов. Следовательно, такие тексты потребовалось проверять еще раз. Этот фактор повлиял на выполнение общей работы, поэтому было принято решение изменить алгоритм работы по подготовке текстов для корпуса калмыцкого языка. Сначала тексты сканировались и автоматически распознавались, затем данные в виде отдельных страниц и изображений сгружались на сайт www.kalmcorpora.ru/reco, к которому имеют доступ только зарегистрированные пользователи. В вычитке текстов приняли участие ученики и учителя средних учебных заведений, студенты Калмыцкого государственного университета, которые с радостью откликнулись на призыв поучаствовать в создании корпуса калмыцкого языка, за что мы выражаем им глубокую признательность и благодарность. Тексты проходили двойную проверку, также на сайте присутствует модуль статистики и отчетности.

На этом этапе мы столкнулись с другой проблемой нетехнического характера: текстов на калмыцком языке ограниченное количество. К тому же очень много переизданий. Например, в сборнике стихов встречается всего три-пять оригинальных текстов, которые ранее или позже не переиздавались. И в конце уже 2013 г. неоцифрованных текстов на калмыцком языке уже стало гораздо меньше, поэтому мы посчитали нужным создать

задел для параллельного подкорпуса текстов, представляющего собой системы выровненных предложений текста-оригинала и текста-перевода.

В ходе реализации проекта были переведены около 20 тыс. текстов на калмыцком языке разной длины. Объем художественных (как прозаических, так и поэтических) текстов составил около 13 млн токенов, газетных — около 7 млн токенов. Это тексты, которые относятся к современному периоду истории языка.

В массиве текстов, подготовленных для корпуса калмыцкого языка, меньшую долю составляют старописьменные памятники на «тодо бичиг», относящиеся в началу XVIII в. Указанные тексты транслитерированы на латиницу, поскольку на данный момент не существует поддержки вертикального письма в текстовых редакторах и отсутствует поддержка UNICODE символов «тодо бичиг». Приводить их к текстовому формату с сохранением графики с использованием неправильно разработанных шрифтов в этот момент бессмысленно, поскольку в базе данных тексты хранятся как линейная последовательность кодировок этих символов, следовательно, отражаться символы в Интернет будут в соответствии с их кодировкой. К тому же выбор транслитерации на латинице обусловлен еще и тем, что в ойратоведении сложилась традиция передачи текста на старокалмыцком письме именно латиницей. Была тщательно исследована система символов «тодо бичиг» UNICODE, создан список символов, не имеющих поддержки UNICODE. Эти данные станут основой технической документации в получении кодировок для некоторых символов в международной организации UNICODE.

Для удаленной работы исполнителей по транслитерации старописьменных текстов создан открытый электронный ресурс по каталогизации текстов на старописьменном языке и их обработке. Поскольку ни одна система не поддерживает вертикального письма (ни Windows, ни Mac) и отсутствует программа распознавания символов «ясного письма», то приходится пока обрабатывать вручную эти материалы. Но это необходимый шаг в этом направлении на первом этапе сбора информации о качестве архивного материала, наборе возможных очертаний символов и т. д. По договору с Национальным архивом Республики Калмыкия были оцифрованы 640 архивных документов, из них транслитерированы 459 листов на старокалмыцком языке, а также обработан в соответствии с принятыми правилами транслитерации памятник старокалмыцкой литературы «Сказание о хождении в Тибетскую страну малодербетовского Бааза-бакши», состоящий из 120 листов, что в целом составило 38 610 слов и

почти 301 000 символов (7,5 п. л.).

В ходе работы над корпусом были написаны несколько вспомогательных программ. Одна из них исправляет кодировку текстов на калмыцком языке, полученных от калмыцкого издательства и редакции национальных газет и журналов, а также транслитерирует по заданным правилам тексты на «тодо бичиг», на латиницу и кириллицу. Вторая сегментирует объемные файлы на части, приписывая уникальный код из базы данных по метаописанию MetaKT, где аннотировано около 18 тыс. произведений и 800 изданий в соответствии со стандартами метаразметки в корпусной лингвистике (Куканова и др., 2012а).

Вышеупомянутая база данных по метаописанию языка структурирована в MS Office Access 2007, позволяющая описывать как тексты, так и их авторов. Кроме того, в ней фиксируется информация, предназначенная для служебного пользования и позволяющая вести учет проделанной работы. База данных состоит из трех взаимосвязанных таблиц, дающие возможность оперативно вносить метатекстовую информацию: «Authors», «Books» и «Texts».

Данная база отвечает за метаописание текстов, т. е. за дополнительную характеристику. По окончании работы по проверке записей и удалении. повторяющихся значений была произведена конвертация файла в формат базы данных MySQL, установленной на сервере, где размещен корпус.

Подготовка материалов для корпуса связана также с записью и сбором образцов устной речи, носящих актуальный характер в свете постепенной утраты языка и уменьшения количества свободно владеющих языком людей по естественным причинам. По характеру собранного материала речь можно классифицировать как публичную, так и непубличную. В качестве публичной речи предоставлены радио- и телезаписи (т. е. аудио- и видеозаписи). Была записана и спонтанная речь во время поездок по районам Республики. Всего 33 часа записи. Произведена первичная обработка звуковых файлов и сегментация. Записанный материал был частично расшифрован, но это всего лишь малая доля из массива собранной устной речи (всего 200 минут в программе ELAN) (более 150 тыс. символов, 28 тыс. токенов). Создана база данных людей, свободно говорящих на калмыцком языке. Разработана архитектура базы данных KalmykSpeech, где дается характеристика информантов и звуковых файлов. Звуковой корпус по калмыцкому языку состоит из трех модулей: базы данных «KalmykSpeech», звуковых файлов и расшифровок в формате .eaf.

Общая метаинформация, собранная в процессе заполнения анкеты, фиксируется в специально созданных таблицах в программе Access MS Office. Таблицы Informants, SoundFiles, Epizods связаны друг с другом при помощи общих полей (ниже приведена схема данных). Этот расшифрованный материал станет заделом для создания устного подкорпуса.

В 2012 г. разработана структура подкорпусов на основе анализа имеющегося текстового материала. На данный момент очевидно, что стоит развивать следующие виды модулей: 1) основной корпус; 2) корпус ранних текстов; 3) диалектный подкорпус; 4) параллельный подкорпус; 5) устный подкорпус; 6) поэтический подкорпус; 7) газетный подкорпус; 8) синтаксический подкорпус; 9) морфемный подкорпус; 10) обучающий подкорпус; 11) фольклорный подкорпус. Отдельным модулем является подкорпус названий. Разработана система морфологической разметки с учетом диалектных особенностей (Куканова и др., 2012б).

К главным подкорпусам, которые были полноценно разработаны, относятся основной (художественный прозаический), поэтический, фольклорный и газетный модули, по которым доступен поиск с учетом калмыцкой морфологии и семантики. Поэтические тексты были подготовлены, по совету д-ра фил. наук С. А. Крылова, специальным образом: мы попытались сохранить деление на строки. В этих целях тексты были автоматически обработаны: конец строки был заменен на символ «/». Тем самым создали маркеры деления на строки, сохранив ритмическую организацию текста, рифму в конце строки и аллитерацию в начале строки.

К концу 2014 г. были созданы заделы для параллельного, старокалмыцкого и устного подкорпусов калмыцкого языка. Как было сказано выше, в ходе реализации проекта был оцифрован массив переводов классической и советской литературы для создания подкорпусов параллельных текстов: русский — калмыцкий и калмыцкий — русский. Для калмыцкого языка параллельные тексты имеют большое значение в аспекте сохранения языка. В процессе изучения калмыцкого языка школьники, студенты и просто желающие изучать язык опираются прежде всего на знания первичного языка — русского, который относится к совершенно другому типу языков (флективный, свободный порядок слов в предложении). Подкорпус параллельных текстов позволит исследователям сопоставить два разноструктурных языка и вывести соответствия на лексико-грамматическом уровне. Реализован поиск по базе выровненных предложений (http://kalmcorpora.ru/parallel). В настоящий момент на сайте Национального корпуса калмыцкого языка размещен пока один

параллельно выровненный текст оригинала и перевода повести А.С. Пушкина «Капитанская дочка», но оцифровка текстов для параллельного подкорпуса продолжается и будет пополняться новыми текстовыми источниками.

Диалектный корпус (http://kalmcorpora.ru/dial) представляет собой уникальное собрание фонетических записей, сделанных Г.И.Рамстедтом в начале XX в. Фиксация калмыцкой речи в соответствии с системой финно-угорской транскрипции передает всю палитру аллофонов калмыцкой речи. Оцифровка и перевод фонетической записи в систему Unicode позволят провести разнообразный анализ представленного диалекта калмыцкого языка в различных аспектах — прежде всего фонетики, лексики и морфологии. Для набора фонетической транскрипции сказок, зафиксированных Г. Рамстедтом, разработана программа Symbol Table, позволяющая работать с заданным списком символов Unicode в различных программах (любые текстовые редакторы).

Старокалмыцкий корпус (http://kalmcorpora.ru/todosearch) — это подкорпус текстов начала XVIII — начала XX в. Оцифрованные и транслитерированные тексты на старокалмыцком языке позволяют описать историю языка. Если развивать проект в дальнейшем (увеличивать его в объемах, снять омонимию и создать разные виды аннотации), то можно получить обширнейший материал для исследования лексики и грамматики. Главной перспективой в технологическом плане, пожалуй, можно назвать создание модуля снятия омонимии.

Проект адресован очень большому кругу пользователей. Например, если студента или школьника интересует, как то или иное слово переводится на калмыцкий язык, он может воспользоваться поиском по переводу. Для исследователя-калмыковеда интерес представляет вся совокупность разработанной разметки как объект лексико-грамматического исследования, а сами тексты как материал исследования с разных точек зрения — в диахронном или синхронном аспекте. Для преподавателей калмыцкого языка материал корпуса можно использовать для разработки упражнений для учеников и студентов в средних и высших учебных заведениях. Преподаватели калмыцкого языка, разрабатывающие основы преподавания калмыцкого языка как иностранного могут найти в созданном ресурсе информацию и сочетаемости слова в лексическом и грамматическом аспекте, что нужно отметить является совершенно не разработанной проблемой в калмыцком языкознании. Корпус предназначен не только для лингвистов, но и для историков и этнографов: тексты содер-

жат уникальную информацию по разным вопросам этнографии и истории, поскольку художественные произведения отражают окружающую реальность, а значит и описания событий, обрядов, традиций и т. п.

Список литературы:

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

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

Дата поступления: 28.11.2014 г.

NATIONAL CORPORA OF THE KALMYK LANGUAGE: RESULTS AND PROSPECTS

V. V. Kukanova

Abstract: The article describes the results of three years work on the project «National Corpus of Kalmyk Language: the creation and development», supported by the Russian Foundation for the Humanities in 2012-2014.

Keywords: corpus linguistics, linguistic databases, computer technology, sysStems architecture of natural language processing, metadescription, tokenization, segmentation, lemmatization, morphological model of language, Kalmyk language.