Тер-Аванесова А. В., Крылов С. А. Использование лексико-грамматических баз данных в русской диалектной лексикографии // Кибрик А. Е. (ред.), Компьютерная лингвистика и интеллектуальные технологии. Вып. 8 (15). По материалам международной конференции "Диалог'2009" (Бекасово, 27 - 31 мая 2009 г.). М.: РГГУ, 2009, с. 471-475.

Тер-Аванесова А. В., Крылов С. А. Лексико-грамматические базы данных и сравнительное изучение русских диалектных акцентных систем // *Кибрик А. Е.* (гл. ред.), Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). Вып. 9 (16). - М.: Изд-во РГГУ, 2010, с. 507-511.

В. В. Куканова, г. Элиста

О НАЦИОНАЛЬНОМ КОРПУСЕ КАЛМЫЦКОГО ЯЗЫКА

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

Как известно, жители Республики Калмыкия постепенно теряют свой родной язык как средство коммуникации и познания, предпочитая осуществлять всю коммуникацию на русском языке, в том числе в семье предпочитают говорить на последнем, что еще негативнее влияет на жизнеспособность языка в обществе. К тому же все процессы, происходящие в языке, «застыли»: без носителей, как правило, язык становится «мертвым».

В этой связи особую актуальность приобретает проект создания и разработки Национального корпуса калмыцкого языка. Создание и функционирование Национального корпуса калмыцкого языка диктуется требованиями времени, инновационной концепцией развития российского общества, прежде всего необходимостью модернизации науки во всех ее областях и повышения ее конкурентоспособности на мировом научном пространстве.

В ходе реализации проекта был проанализирован опыт отечественной и зарубежной лингвистики по созданию национальных корпусов, объемных текстотек (преимущественно языков агглютинативной структуры – башкирского, татарского и др.). В первую очередь внимание уделялось вопросам репрезентативности текстового материала в аспекте проблем исчезающих языков. Текстов на калмыцком языке достаточно мало, объемы 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). Для увеличения корпуса было принято решение оцифровать все, что имеется, конечно, при этом определяя приоритетность источников в сканировании и распознавании.

Работы по созданию Национального корпуса калмыцкого языка ведутся с конца 2010 года. В начале 2011 года Калмыцкий институт гуманитарных исследований Российской академии наук обратился за помощью в Издательский дом «Герел» и в редакцию газеты «Хальмг Үнн» в сборе текстов на калмыцком языке, и они, конечно же, любезно предоставили свои электронные архивы, что стало заделом в создании корпуса калмыцкого языка. Несмотря на это, тем не менее, мы столкнулись с проблемой: все тексты, которые содержатся в программах, предназначенных для верстки (Indesign, PageMaker), с нарушенной кодировкой. Буквы калмыцкого алфавита, не совпадающие с кириллицей, не имеют кодировки ANSI, но их поддерживает UNICODE, но, поскольку шрифты для калмыцкого языка создавались непрофессионалами, то графемы и глифы имеют кодировку совсем других символов. Для решения этой проблемы была написана программа, исправляющая кодировку текстов на калмыцком языке, полученных из издательства и редакции национальных газеты и журнала*. Общий объем электронного архива составил около 7 млн словоупотреблений.

Отдельно велась оцифровка текстов: сначала тексты сканировались в программе Abby FineReader 11 Pro, которая поддерживает распознавание калмыцкого языка, однако без словарной поддержки. Для улучшения распознавания мы воспользовались режимом обучения и созданием пользовательского словаря, что значительно улучшило качество распознавания. Тем не менее ошибки все же остаются, вычитка распознанных текстов позволила выявить следующие типы ошибок: 1) *у* часто распознается как *у*: *куукн* вместо *күүкн* ‘девочка’; 2) *е* при плохом качестве оригинала распознается как *с*; 3) *җ* распознается как *ж*; 4) *ң* как *н*; 5) *һ* как *h*, т. е. программа не различает написание прописной и строчной графемы и т. д. Но в принципе ошибок не так много, пользовательский словарь существенно уменьшает их количество. Например, если слово *күн* ‘человек’ часто распознается без словарной поддержки как *кун* при плохом качестве оригинала (недостаточно четко пропечатанных букв), то при словарной поддержке, созданной для Abby

* Данная программа транслитерирует по заданным правилам тексты на «тодо бичиг», на латиницу и кириллицу.

FineReader 11 Pro в пользовательском режиме, такие отрезки распознаются правильно, поскольку в калмыцком языке и соответственно в словаре отсутствует слово *кун*.

В 2012 году распознавание велось в программе Abby FineReader 11 Pr, но возникла проблема при конвертации проверенных текстов в текстовые форматы: сохраняются ошибки, наличие ошибок с переносами, следовательно, такие тексты требовалось проверять еще раз. Поэтому было принято решение изменить алгоритм работы над оцифровкой текстов. Сначала тексты сканируются и распознаются автоматически, затем данные в виде отдельных страниц и изображений сгружаются на сайт www.kalmcorpora.ru/reco, где только зарегистрированные пользователи имеют доступ к вычитке текстов (см. рис. 1). В этой работе принимают участие средние учебные заведения, которые с радостью откликнулись на призыв поучаствовать в создании корпуса калмыцкого языка, за что мы выражаем им глубокую признательность и благодарность. Тексты проходят двойную проверку, также на сайте присутствует модуль статистики и отчетности. На данный момент мы оцифровали текстов объемом около 5 млн словоупотреблений, и работа по их проверке продолжается. Однако 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. Сейчас мы буквально выискиваем новые тексты (например, в журнале «Теегин герл»), а на это уходит очень много времени.

Выше описаны два направления в создании корпуса калмыцкого языка. Отдельно в рамках проекта мы собираем материал по старописьменному языку – старописьменные памятники на «тодо бичиг». Был частично оцифрован знаменитый фонд И-36 архива Национального архива Республики Калмыкия, содержащий более 400 дел*. Последние были транслитерированы на латиницу, поскольку на данный момент не существует поддержки вертикального письма в текстовых редакторах, а также это соответствует традиции изучения текстов на старокалмыцком языке. Была тщательно исследована система символов «тодо бичиг» UNICODE, в результате чего создан список графем и глифов, не имеющих поддержки UNICODE. Сейчас находится в стадии разработки модуль по каталогизации текстов на старописьменном языке и их обработке, поскольку ни одна система не поддерживает вертикальное письмо (ни Windows, ни Mac) и отсутствует программа распознавания символов «ясного письма», то приходится пока вручную обрабатывать эти материалы. Но это необходимый шаг в этом направлении, так как, думается, это первый этап сбора информации о качестве архивного материала, наборе возможных очертаний символов и т.д.

Конечно же, вся эта работа сопровождается метаописанием материала, сконструированным в программе 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 [Куканова, Бембеев, Мулаева, Очирова 2012].

Разработана структура подкорпусов на основе анализа имеющегося 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 [Куканова, Бембеев, Мулаева, Очирова 2012]. В рамках проекта полноценно развиваться будет три модуля: газетный, основной (проза разных стилей) и поэтический (но без стихотворной разметки).

Составлен грамматический словарь калмыцкого языка на 27 тыс. единиц** на основе словника калмыцко-русского словаря под ред. Б. В. Муниева [Калмыцко-русский словарь 1977; Куканова 2012а; 2012б; 2013]. Разработана система морфологической разметки с учетом диалектных особенностей [Куканова, Очирова 2012], приняты решения по созданию фольклорного подкорпуса.

Для нас, жителей Республики Калмыкия, особенно значима практическая значимость этого проекта, в первую очередь его ценность заключается в том, что корпус и результаты корпусных исследований можно будет использовать в обучении детей и студентов калмыцкому языку. Во-первых, материалы корпуса можно будет применять на уроках калмыцкого языка как средство составления различных «простых» упражнений и заданий (расставить знаки пунктуации, вставить буквы, найти однокоренные слова и многое другое). Можно ставить перед учениками и вопросы

---

* Оцифрованы только документы на «тодо бичиг» в виде отсканированных изображений.

** Данная работа введется в рамках темы НИР, и она необходима была для создания морфологического анализатора.

исследовательского (эвристического) характера (например, найти функциональные особенности разных стилей на основе сравнения двух текстов и т. п.). И это только взгляд с одной стороны и при этом самый простой способ применения возможностей корпуса.

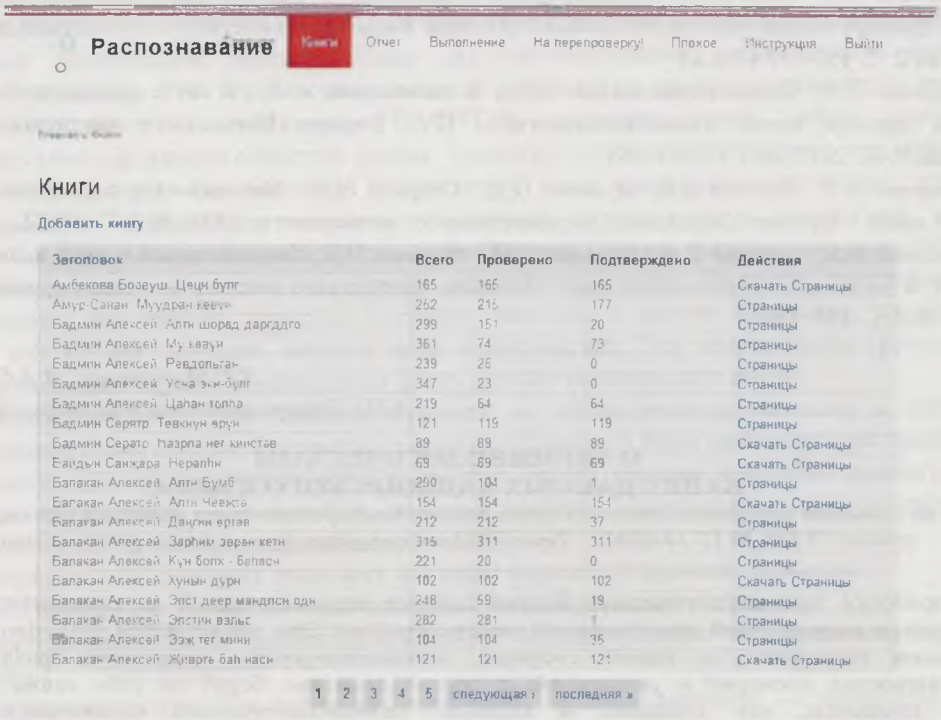

Рис. 1. Портал по проверке текстов на калмыцком языке.

С другой стороны, результаты исследований, основанных на корпусных методах, помогут скорректировать имеющиеся программы изучения калмыцкого языка и даже создать новые, которые будут учитывать реальное функционирование языковой системы. А. Ш. Кичиков не раз предлагал строить образовательные стандарты исходя из методики обучения языку как иностранному с введением уровневой системы знаний языка (в соответствии с международными классификациями) и с учетом международного опыта по изучению английского языка, поскольку здесь методы и приемы по сути одинаковы, нужна лишь небольшая адаптация под специфику нашего родного языка. На основе корпуса можно провести исследования частотности калмыцких слов, конструкций, создать словарь сочетаемости лексических единиц, дать детальное описание грамматики в том виде, в котором она реально функционирует, и многое другое, что действительно необходимо для сохранения языка, ведь это один из «работающих» способов ревитализизации языка в калмыцком обществе.

Таким образом, работа по созданию и развитию Национального корпуса калмыцкого языка ведется достаточно активно, хотя на этом пути были и ошибки, чего было сложно избежать, поскольку для нас это первый опыт работы по созданию подобных ресурсов. В настоящем виде описанный алгоритм работы может изменяться, дополняться, совершенствоваться в зависимости от ситуации и условий работы.

Литература

Калмыцко-русский словарь / под ред. Б. Д. Муниева. М.: Изд-во «Русский язык», 1977. 768 с.

Куканова В.В. Морфологическая модель калмыцкого языка в свете автоматической обработки текстов: предварительные замечания // Информационные технологии и письменное наследие. El'Manuscript-2012: Материалы IV международной научной конференции (Петрозаводск, 3–8 сентября 2012 г.). Петрозаводск, Ижевск, 2012. С. 142–146.

Куканова В.В. Словоизменительные глагольные типы в калмыцком языке в свете автоматической обработки текстов // Вестник Калмыцкого государственного университета. 2013. № 3. (в печати)

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. (1 п.л.)

Куканова В.В. Словоизменительные типы в калмыцком языке в свете автоматической обработки текстов (на примере имени существительного) – II // Вестник Калмыцкого института гуманитарных исследований РАН. 2012. № 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. 2012. № 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. 2012. № 3. С. 138–150.

Л.Г.Мигранова, З.А.Сиразитдинов,
А.Ш. Ишмухаметова, А.Д. Ибрагимова, , г. Уфа

О ТЕРМИНОЛОГИЧЕСКОМ БАНКЕ ДАННЫХ БАШКИРСКОГО ЯЗЫКА

Работа по созданию терминологического банка данных башкирского языка ведется в рамках реализации гранта РГНФ № 12-14-02021 "Терминологическая база данных башкирского языка".

Разработка терминологических банков данных является одной из приоритетных и бурно развивающихся направлений практической лексикографии. Это объясняется тем, что базы данных таких банков позволяют, с одной стороны, автоматизировать подготовку различных видов терминологических словарей и указателей, с другой стороны, берут на себя такие трудоемкие и рутинные процессы, как создание и ведение терминологических справочных картотек и распространение самой терминологической информации в отраслях. Отметим, что терминологические базы данных не только значительно расширяют существующие возможности вышеуказанных работ, но и позволяют:

• тиражировать в необходимом количестве однажды введенную информацию или ее часть;

• оперативно выдавать специалистам различные виды информации о терминах в самых разнообразных комбинациях на основе различных сочетаний признаков;

• обмениваться информацией с другими банками данных, что значительно расширяет объем ввода информации без необходимости ее трудоемкой предварительной разработки;

• собирать и хранить огромное количество информации, что дает уникальную возможность инвентаризации максимального количества информации о терминологии с привлечением разнообразных источников, отражающих состояние специальной лексики на разных этапах развития языка;

• быстро получать самые различные статистические данные [Гринев-Гриневич 209, 129].

В мире существует достаточное количество уже разработаннные крупных терминологических банков, которые активно используются в качестве автоматических словарей — переводных, толковых, информационных. Таковыми являются LEXIS (Федеральное бюро языков, Германия), TERMIUM (Оттава, Канада), EURODICAUTON (Комиссия европейских сообществ, Люксембург), Британский лингвистический банк данных [Багаряцкая, Казакевич, Павлович, Членова 1988, 28]. База данных РОСТЕРМ (Российский национальный банк данных по стандартизованной терминологии) [www.gostinfo.ru/show.php?/inf_res_8/r8.htm], Терминологическая база знаний «Научная терминология» (Терминологический центр Института русского языка им. В. В. Виноградова РАН) [www.ruslang.ru/agens.php?id=terminol_centre].

Отметим, банк данных РОСТЕРМ является самым полным отечественным банком, содержит свыше 140 тысяч терминологических статей из ГОСТ, ГОСТ Р, стандартов ИСО и МЭК, а также терминологических приложений к ним, является коммерческим продуктом.

В башкирской лексикографии существуют десятки двуязычных терминологических словарей по многим направлениям науки и техники, однако единого банка данных пока нет. Следует отметить, что некоторые словари уже устарели и требуют пересмотра, дополнения новыми терминами. Создание единого банка данных позволило бы оперативно решить и эту проблему.