УДК 81'33
ББК 81.23

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

МЕТАРАЗМЕТКА В НАЦИОНАЛЬНОМ КОРПУСЕ КАЛМЫЦКОГО ЯЗЫКА

*Статья подготовлена в рамках проекта РГНФ «Национальный корпус калмыцкого языка» (12-04-12047, тип «в»).

Аннотация. *В статье посвящена описанию базы данных, которая включает информацию о текстах в Национальном корпусе калмыцкого языка. Структура базы данных состоит из нескольких взаимосвязанных таблиц, описывающих как тексты, так их авторов. Рассматривается и возможность использования базы данных автономно, т. е. вне корпуса калмыцкого языка.*

Ключевые слова: *корпусная лингвистика, Национальный корпус калмыцкого языка, метаразметка, база данных, корпус названий.*

V.V. Kukanova, E.V. Bembeev, N.M. Mulaeva, N.Ch. Ochirova

METAANNOTATION IN THE NATIONAL CORPORA OF KALMYK LANGUAGE

Annotation. *The article is dedicated to the description of base which includes the information about texts in the National corpora of Kalmyk Language. The structure of base consists of the some tables described texts and authors as well. There is considered a possibility of usage of base independently, outside corpora of Kalmyk language.*

Key words: *corpora liguistics, National corpora of Kalmyk Language, metarazmetka, corpora of name.*

Появление компьютерной науки в середине XX в. и ее последующее развитие до настоящего времени способствовали зарождению вычислительной лингвистики, поскольку любой программный код должен иметь средство и способ коммуникации, другими словами язык, при помощи которого могут передаваться те или иные сообщения. Первый корпус появился почти сразу же, как только программисты стали использовать достижения своей науки в гуманитарной сфере.

Язык долгое время изучался на основе материала, полученного путем ручного фиксирования отдельных примеров из небольшого количества текстов. При исследовании той или иной темы большое время уделялось именно сбору материала, его фиксированию на карточках и затем его обработке. Как правило, в ходе исследования создавалась относительно большая картотека изолированных примеров, например слов и предложений. Если лингвисту необходимо было расширить контекст, то приходилось заново открывать текст и анализировать данный пример в уже большем окружении, что создавало неудобства. Конечно, все это сказывалось на скорости и качестве выполнения исследования. Известно, что человек в силу разных обстоятельств может не заметить тот или иной пример, а иногда и самый главный.

Лингвистический же корпус представляет собой также картотеку, достаточно своеобразную, где зафиксировано и морфологически (иногда семантически) разобрано

каждое слово из текста: ср., «понятие корпуса является продолжением традиционных картотек, с которыми всегда работали лингвисты» [1]. Тем самым исследователь имеет полный спектр функционирования языковых фактов.

Период конца 1980-х и 1990-е гг. отмечены массовым появлением корпусных исследований на материале английского, итальянского, финского и ряда других языков (преимущественно, европейских). Сегодня текстовые корпусы – это мощные информационные ресурсы, которые могут быть использованы в различных исследованиях, прежде всего в лексикографии [2].

Текстовые корпусы на том или ином языке необходимо создавать по нескольким причинам: во-первых, поиск материала для исследования происходит в реальном времени и доступен каждому, у кого компьютер имеет выход в Интернет; во-вторых, существует возможность многократного использования в различных аспектах всего того, что обработано компетентным лингвистом всего лишь раз; в-третьих, поиск необходимых единиц происходит в считанные секунды.

Национальный корпус калмыцкого языка также представляет собой электронную коллекцию размеченных текстов, написанных на калмыцком языке. Во множестве текстов, которые готовятся для корпуса (исправляется кодировка символов, редактируются и форматируются сами тексты), легко запутаться, поскольку они принадлежат разным стилям, разным жанрам, разным вариантам и формам языка. Как среди такого подмножества лингвисту выбрать то, что необходимо для исследования?

В корпусной лингвистике уже есть выработанная схема решения данной проблемы. Это введение метатекстовой информации, которая характеризует текст в целом. Наличие такой информации позволит исследователю очертить рамки поисков языковых фактов и явлений в многомиллионном массиве текстов. Признаки, которые необходимо описать в данном модуле, должны быть релевантными, то есть те, которые могут влиять на характеристики текста. Например, с одной стороны, время создания текста обусловливает особенности языка автора, с другой – время, описываемое в тексте, отражает особенности эпохи (автор, используя прием стилизации, вводит названия предметов и реалий в лексическую систему произведения, которые отсутствуют уже во время создания текста).

В настоящей статье описывается система метаразметки, принятая в Национальном корпусе калмыцкого языка. Несомненно, что тексты на любом языке должны сопровождаться метатекстовой информацией. Под матеразметкой мы понимаем систему помет экстралингвистического характера, или внешнего аннотирования, а также информацию технической работы с текстом (т. е. служебную).

Существует несколько лингвистических программ, которые позволяют проводить метаописание текстов. Это прежде всего Systematic Coder [6] и UAM Corpus [5]. Данные программы находятся в открытом доступе и являются бесплатными, поэтому каждый желающий может просмотреть или даже модернизировать исходные коды. Все они соответствуют стандартам, принятым в корпусной лингвистике. На их основе была выработана архитектура метаописания текстов на калмыцком языке [3].

Однако уже на практике многие ее аспекты оказались лишними, достаточно сложными, в некоторой степени избыточными в разметке массива текстов. К тому же, по замечанию А.Е. Полякова, «полные правила TEI очень обширны и не всегда мотивированы, поэтому соблюсти все требования стандарта достаточно трудно. Формат не отличается компактностью, поэтому часто разметка разрастается без увеличения содержательной информации» [4].

В ходе обработки текстов было принято решение хранить метатекстовую информацию текстов корпуса калмыцкого языка в базе данных, сконструированную в MS

Access 2007. База данных состоит из трех взаимосвязанных таблиц, которые позволяют оперативнее вносить метатекстовую информацию: «Authors», «Books» и «Texts».

Таблица «Authors» включает характеристику авторов. Здесь фиксируется ФИО автора на русском и калмыцком языках, а также его псевдоним, известный широкому кругу читателей. В таблицу входят записи «автор не установлен» и «коллективный автор». Первое поле указывается в случае, когда автор не известен, и чаще всего это группа авторов, которые создают один продукт, например конституцию. Если же авторство текста принадлежит всему народу, то здесь используется в описании вторая вышеуказанная запись. Авторов, пишущих на калмыцком языке, не так много, как, например, на русском языке. Все наиболее известные калмыцкие поэты и писатели зафиксированы в таблице, на данный момент количество записей превышает 230, в их число входят и переводные авторы, например А.С. Пушкин, И.С. Тургенев, А.П. Чехов и другие, а также ряд авторов, создавших свои произведения на старокалмыцком языке. Надо сказать, что таблица постоянно пополняется новыми персоналиями, поскольку калмыцкая литература насчитывает всего несколько столетий, большая часть из которых принадлежит периоду, когда произведения писались на старой письменности («тодо бичиг»). Наиболее известные из них уже зарегистрированы (например Зая-пандита), и теперь остались те авторы, которые не известны широкому кругу читателей или их произведения были обнаружены сравнительно недавно и до сих пор еще не введены в научный оборот.

В данной таблице дается также более детальная характеристика автора:

– годы жизни (годы рождения и смерти, если это возможно);

– пол (муж. и жен.);

– место рождения;

– первый язык (указывается язык, который был первичным в коммуникации и познании);

– диалект (три пометы: “torgut”, “buzav”, “derbet”);

– знание других языков;

– уровень образования;

– род занятий.

В некоторых случаях невозможно дать исчерпывающую характеристику того или иного автора в силу того, что о нем мало известно или до нас дошли отрывочные сведения о его биографии. Однако по возможности нужно давать полную информацию, поскольку эти атрибуты являются релевантными в изучении влияния социологических факторов на порождение речи. Конечно, все факторы невозможно учесть в процессе исследования, так как их очень много. Это могут быть и ситуативные причины, о которых в большинстве случаев мы ничего не знаем. Но, тем не менее, наиболее адекватные мы постарались указать.

Вторая таблица «Books» описывает книги и в основном служебного характера. Каждая книга имеет свой уникальный номер (тип поля: счетчик), здесь же помечаются редакторы, как правило, ими выступали сами поэты и писатели. Отдельное поле Переводчик (Translator) необходимо для параллельного подкорпуса. Отдельное поле Выходные данные фиксирует библиографическое описание, при этом было решено не разделять каждый его элемент в отдельное поле, а объединить все библиографическое описание в одно поле. Далее идут служебные поля:

– источник получения текста (издательство, сканирование, электронная копия, ручной набор, Интернет);

– исполнители.

Такая таблица очень удобна тем, что в калмыцкой книжной литературе XX века достаточно много переизданий, которые полностью повторяют предыдущие книги, являются результатом контаминации нескольких сборников или только часть (значительная или незначительная по своему объему) является оригинальной. Когда тексты уже зафиксированы по одному сборнику, то внести повторяющееся название можно только после сравнения двух произведений. Столбец, где указывается название текста, является индексируемым полем, при этом повторения недопустимы: система предупреждает, что сохранить данную запись невозможно, поскольку данная запись уже содержится в базе. Однако у авторов существуют произведения с одинаковыми названиями, поэтому после предупреждения, которое выдает программа, необходимо присвоить названию дополнительный индекс, если это все-таки разные произведения. В результате внесения записей в базу данных можно понять, какие книги нам нужно сканировать, а какие не нужно, поскольку они являются переизданиями материала предыдущих книг. Из них выбираются те, которые обладают высоким качеством, необходимым для ускорения процесса распознавания.

Каждому тексту присвоен свой уникальный номер, это ключевое индексируемое поле, создающееся автоматически (тип поля: счетчик). Этот Id позволит связать размеченный с морфологической точки зрения текст и информацию о нем.

В поле «Название текста» фиксируется заглавие текста, вплоть до стихотворений, которые принадлежат жанру «ахр шүлгүд», сотоящие, как правило, из двух-четырех строк. Если название отсутствует, то фиксируется первая строка текста. В полях «Автор» и «Переводчик» из выпадающего списка выбирается автор текста и переводчик, если текст принадлежит переводной литературе.

В поле <lang> указывается язык, на котором впервые создан текст. В большинстве случаев это «xal», однако в некоторых случаях, в частности для параллельного подкорпуса, отмечается оригинальный язык, на котором создано то или иное произведение. В процессе обработки книжных фондов на калмыцком языке выяснилось, что у нас имеется достаточно большое количество переводных текстов, преимущественно классической литературы с русского языка.

Следующие поля включают уже описание атрибутов самого текста:

– форма речи: письменная или устная;
– графическая система;
– тип речи: поэзия или проза;
– стиль текста: художественный, официально-деловой, научный, разговорный и т.д.;
– жанр текста, включает большое количество элементов;
– выходные данные;
– страницы;
– исполнитель;
– дата регистрации;
– комментарий.

Образцы таблиц представлены ниже (см. рис. 1 и 2). База данных является способом репрезентации метаинформации. Надо сказать, что в процессе заполнения таблиц выявилось, что подавляющее большинство текстов на калмыцком языке – это тексты поэтические. Прозаических текстов достаточно мало, а, как известно, именно они составляют каркас корпуса. Изучать свойства языка на поэтических текстах весьма сложно, поскольку текст стихотворения подчиняется другим правилам. Слово здесь может быть употреблено в метафорическом или метонимическом значении. Создание поэтического языка специфично для каждого автора, и выделить универсальное и специфическое в языке сложно.

База данных перепроверяется исполнителями, устраняются ошибки. Перед публикацией на сайте материалов база данных объединяется с размеченным текстом. Надо сказать, что база данных может быть использована автономно в различных научных целях. Например, созданная база данных может послужить материалом для корпуса названий текстов. Объем базы уже репрезентативен для проведения исследований по заголовкам текстов, написанных на калмыцком языке, при этом она постоянно пополняется новыми записями. Базу данных можно использовать и для поиска материала, где реализована одна и та же тема. Она поможет проследить изменения как хронологически, так и по авторам.

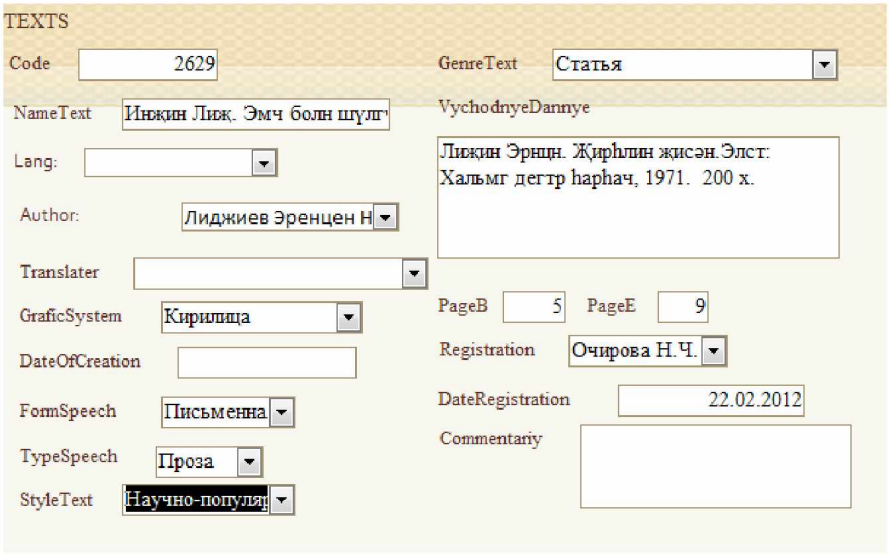

Рис. 1. *Образец формы таблицы «Text»*

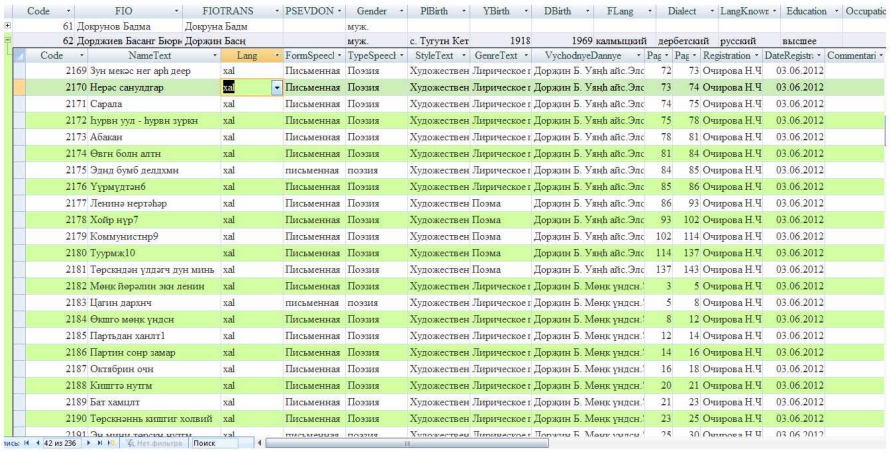

Рис. 2. *Образец таблиц «Authors» и «Text»*

Список литературы

1. Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. – Иркутск: ИГЛУ, 2005. – 161 с.

2. Крылов С.А. Стратегии применения интегрированной информационной среды StarLing в корпусной лингвистике и в компьютерной лексикографии // Ред. И.С. Смирнов. Orientalia et classica. Труды Института восточных культур и античности. – Вып. XIX. Аспекты компаративистики. – М.: РГГУ, 2008. – С. 649–668.

3. Куканова В.В. Архитектура метаописания в Национальном корпусе калмыцкого языка // Вестник Калмыцкого института гуманитарных исследований РАН. – 2011. № 1. – С. 139–145.

4. Поляков А.Е. Технология подготовки информации в Национальном корпусе русского языка // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. – М., 2005. – С. 175–192.

5. UAM Corpus [электронный ресурс] // http://www.wagsoft.com/Coder/ (дата обращения: 15.07.2012).

6. Systematic Coder – a Text Markup Tool [электронный ресурс] // http://www.wagsoft.com/CorpusTool/ (дата обращения: 15.07.2012).