ФОЛЬКЛОРНЫЙ ПОДКОРПУС: ПРОБЛЕМЫ, СТРУКТУРА И ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ*

В. В. Куканова

Языковая система, ее структура, функционирование, взаимосвязи ее элементов всегда изучались на материале письменных или устных источников, где языковые факты находят свое выражение, и не случайно, что в последние 50 лет в связи с достижениями в области вычислительной техники популярным направлением в лингвистике стали корпусные исследования, которые основаны на текстах, образующих корпус. В языкознании сложились два общепринятых значения термина *корпус*. Различают соответственно значениям «корпус первого порядка», под которым понимают коллекцию текстов на некотором языке, и «лингвистический корпус» (языковой). Последнее более объемное и системное понятие, называющее ряд признаков-критериев, которыми должна обладать то или иное собрание текстов. Это «…большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [Захаров 2005: 4]. Итак, коллекция текстов может быть названа корпусом, если она обладает репрезентативным объемом, представлена в электронном аннотированном виде как определенная система.

Перед языковедами, изучающими калмыцкий язык, также стоит задача создания подобной системы, она более актуальна в данный момент, чем задачи исследования, поскольку калмыцкий язык находится на грани исчезновения и входит в список, согласно терминологии ЮНЕСКО, «definitely endangered language» («определенно исчезающих языков»). К этой группе относятся языки, не все носители которых говорят на родном языке. Передача знаний о языке от старшего поколения к младшему утрачена, в семье также не говорят на родном языке, предпочитая другой более престижный в социальном плане язык [ЮНЕСКО].

В создании, а затем и развитии данного электронного ресурса лингвисты встречают множество проблем как теоретического, так и практического планов. К теоретическим проблемам следует отнести неразработанность морфологической структуры языка, спорные вопросы, касающиеся частеречной принадлежности лексических единиц, дифференциации фраз и сложных слов и многое другое. К практическим вопросам можно отнести нарушение кодировки символов и глифов в используемых шрифтах, главным образом у специфических букв калмыцкого алфавита. Существует проблема репрезентативности текстового материала, поскольку для более глубокого и системного описания, например, грамматических категорий необходимо иметь достаточно объемный ресурс не менее 100 млн словоупотреблений, что уже, на наш взгляд, сделать почти невозможно или весьма трудоемко. Все это во многом усложняет и определенно задерживает работу по созданию корпуса.

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

В структуре будущего корпуса предварительно можно будет выделить следующим образом:

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;
12)подкорпус названий.

Реализация каждого подкорпуса требует множества предварительных шагов, создания словарных материалов, подготовки текстов, разработки концепции каждого подкорпуса, аннотации и разметки. В данной статье мы остановимся на фольклорном подкорпусе калмыцкого языка, поскольку необходимо предварительно осветить проблемы, связанные с его созданием, и попытаться найти определенные решения этих проблем, аргументируя при этом свою позицию.

Что касается фольклорных баз данных, то следует отметить, что только в последнее время стали задумываться об их создании. Подобных проектов немного, что обусловлено, с одной стороны, отсутствием поддержки со стороны государства, с другой – отсутствием единой системы хранения фольклорного наследия, единого алгоритма действий, отсутствием технических специалистов для сохранения и оцифровки фольклорных данных, слабой технической материальной базой Институтов и архивов.

Фольклорные произведения являются неотъемлемой составляющей духовного наследия любого народа, в частности и калмыцкого этноса. 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), как нам кажется, необходимо включить фольклорные произведения в Национальный корпус калмыцкого языка. Оцифрованная база данных позволит сохранить фольклорное наследие калмыцкого этноса для последующего изучения как лингвистами, так и фольклористами, и этнографами. Отсюда и вытекает одна из проблем – проблема адресата, поскольку именно он определяет особенности подготовки текста и структуры корпуса.

Об этом пишет, например, и Д. С. Лихачев: «Следует прежде всего различать издания для лингвистов и издания для литературоведов и историков. Лингвист кладет в основу издаваемого текста список, наиболее ценный с точки зрения языка. Литературоведы и историки кладут в основу издания список, дающий наилучшее представление о памятнике или о редакции памятника. Лингвисты нуждаются по преимуществу в издании текста одного списка. Литературоведы и историки стремятся издавать текст по всем спискам, дать представление о редакциях, об истории текста памятника. Различается у лингвистов, с одной стороны, и у литературоведов с историками — с другой, выбор разночтений. У лингвистов разночтения должны

давать представление о языке списков; у литературоведов и историков — об изменениях в содержании в отдельных списках, а специально у литературоведов — и об изменении стиля» [Лихачев 1964: 69].

Именно адресат создаваемого подкорпуа обуславливает остальные проблемы: если лингвисту достаточно одного текста - авантекста, который, на его взгляд, содержит интересные языковые факты, то для фольклориста важно иметь разные версии того или иного произведения. Вариативность фольклорных произведений – это обязательный их признак. Степень их различия может быть различной по содержанию, по форме, по степени сохранности: от сильной до незначительной. Во-первых, варианты могут отличаться друг от друга по содержанию. Во-вторых, они могут дифференцироваться по форме, при этом говорить можно о легкой трансформации текста, когда заменяются словами, близкими по значению, синонимичными конструкциями, выражая один и тот же смысл, или сильной степени трансформации текста, когда варианты отличаются друг от друга художественной формой, например, один вариант существует в прозаической форме, другой – в поэтической. В-третьих, тексты инвариантов фольклорного произведения могут отличаться степенью сохранности. Конечно, существуют их возможные комбинации: содержание и форма одинаковы, но некоторые части текста утрачены; содержание и сохранность не изменяемы, а форма инвариантов отличается; форма и сохранность одинаковы, а содержание разное и т. д. Именно подобные изменения в ткани текста интересны фольклористам, поскольку в ходе анализа вариантов того или иного произведения фольклора можно понять причины и смысл этих изменений. Исследование вариантов одного и того же текста позволяет выявить закономерности развития фольклора.

Однако и текстах вариантов фольклорных произведений лингвист можно найти материал для исследования (языковые факты), думается, не следует замыкаться только на исследовании грамматических явлений, поскольку такие тексты могут содержать и диалектные, и орфоэпические особенности, такие тексты интересны и для психолингвистов, изучающих порождение речи, и для типологов, которые изучают универсальное и специфическое на материале одного или нескольких языков.

К этой же группе проблем относится вопрос фиксирования, вернее отражения в корпусе разновременных записей. Они также позволяют проследить динамику развития фольклорного наследия этноса. Думается, что подобные тексты будут содержать языковые особенности своего времени. Лингвисты могут проследить диахронию и синхронию языкового среза, отраженного в одном конкретном произведении.

Для лингвистов в большинстве случаев достаточно небольшого отрезка текста, максимум – абзац (оговоримся, что речь не идет о лингвистах, изучающих высшую единицу языка – текст – и особенности диалектов и подговоров), чтобы изучить системно то или иное лингвистическое явление, грамматическую категорию и т. п. Для фольклористов же удобнее по запросу получать целый текст или более цельный его фрагмент, позволяющий анализировать сюжеты и мотивы фольклорных произведений, их композиционную структуру и непосредственно их язык, но не просто язык, а языковые особенности и функции на фоне всего текста. В данном случае, видимо, в корпусе, размещенном в Интернете, будет размещена версия, в которой по запросу можно будет получить только абзац, полнотекстовое собрание фольклорных текстов будет доступно сотрудникам Калмыцкого института гуманитарных исследований РАН и может распространяться на дисках.

К следующему вопросу, который требует своего решения, можно отнести текстологические проблемы. Известно, что ряд ранее опубликованных текстов готовился без учета текстологических принципов, например С. Н. Азбелева [1966], К. В. Гацака [1989] и К. В. Чистова [2005]. Эти источники являются результатом контаминации разновременных записей, редактирования текста. Встает вопрос о правомерности включения подобных текстов в подкорпус, поскольку в настоящее время в Калмыцком институте гуманитарных исследований РАН ведутся работы по изданию фольклорных произведений в соответствии с современными текстологическими принципами. Одним из главных этих принципов является подготовка текстов в оригинальном виде «с сохранением всех особенностей того или иного сказителя (например, индивидуального произношения или диалектизмов в тексте): „…мы обязаны признавать основной ту из нескольких форм одной записи, которая в наибольшей степени отражает конкретный народный вариант, зафиксированный собирателем“ [Гацак 1971: 111]» [Убушиева 2011а: 168]. Фольклористы Института пытаются восстановить оригинальные записи фольклорных произведений, привести разные версии, которые отличаются друг от друга временем записи, исполнителем, степенью сохранности и др. Текстологическая работа по сличению разных текстов фольклорных произведений уже выявила факты контаминации, пропусков, замен, добавлений строк и слов, изменение их форм (см., например, работы [Убушиева 2011а; 2011б; Манджиева 2011а; 2011б]]).

В связи с вышесказанным резонным является вопрос о включении текстов фольклорных произведений, ранее изданных и противоречащих теоретическим основам текстологии. Представляется, чтобы избежать дублирования текстов, не совсем соответствующих оригиналу, нужно обрабатывать эти тексты только после того, как они будут опубликованы в своем новом варианте. Однако уже на этот момент существует несколько изданий фольклорных произведений, которые были подготовлены в соответствии с требованиями современной эдиции. Например, Сказки Санджи Бутаева (Буутан Санҗин туульс) [2008], Хранитель мудрости народной – Ш. Боктаев (Алтн чеежтә келмрч Боктан Шаня) [2008], Калмыцкие народные благопожелания (Хальмг улсин йөрәлмүд) [2010], Фольклорные материалы из репертуара Т. С. Тягиновой (Т. С. Тягинован амн урн үгин көрңгәс) [2011] и др.

Структура фольклорного подкорпуса имеет общую метаразметку (см. [Куканова 2011]) и свою собственную, которая более детализирована и учитывает сведения, необходимые для фольклорных и этнографических исследований. Метаразметка в фольклорном подкорпусе имеет следующую структуру:

1) идентификационный код текста;
2) название текста;
3) перевод названия на русский язык;
4) исполнитель/сказитель;
   - годы жизни;
   - родовая принадлежность;
   - социальное положение;
5) ФИО человека, зафиксировавшего текст (собиратель);
   - годы жизни;
   - национальность;
   - род занятий;
6) форма фиксации (цифровая/аналоговая запись или письменный текст);
7) время фиксации;

8) место фиксации;
9) место хранения оригинала или записи;
10) шифр;
11) количество листов;
12) наличие копии;
13) внешняя сохранность/качество записи;
14) датировочные данные;
15) сведения о почерке;
16) история публикации;
17) графическая система;
18) жанровая принадлежность;
19) комментарий.

Метаописание текста дает краткие сведения об археографии и текстологии документа, по которым исследователь может дать общую информацию о фольклорном тексте.

Первичная обработка фольклорного материала заключается в предварительном метаописании текста и его орфографическая расшифровка или транслитерация с «тодо бичиг» на латиницу и современный калмыцкий язык. Такая работа замедлит, с одной стороны, процесс, с другой – увеличит возможности использования фольклорных материалов, ведь на их основе можно провести качественное лингвистическое, текстологическое, фольклорное исследования.

Фольклорный подкорпус состоит из двух модулей: массива оригиналов (в виде отсканированных документов или в виде записей) и базы данных KalmFolkTexts. Первый массив – это преимущественно файлы формата .jpg и аналоговые записи (отметим, что в рамках другого проекта, реализуемого в Институте ведется работа по конвертированию звуковых файлов в цифровой). Последняя представляет собой реляционную базу данных, разработанную в формате MS Office Access. На настоящий момент она состоит из 6 таблиц, которые условно можно поделить на две группы: фактические данные (о фольклорном произведении, исполнителе) и результаты научно-исследовательской работы и их интерпретация. Некоторые таблицы содержат «смешанные» данные. Структура базы данных следующая:

- Таблица «Исполнитель»;
- Таблица «Собиратель»;
- Таблица «Текст»;
- Таблица «Предложения»;
- Таблица «Слова»;
- Таблицы «Глоссы».

В таблице «Текст» имеется и служебная информация: кто расшифровал/транслитерировал, кто проверял/перепроверял. Двойная проверка обоснована тем, что транслитерация и расшифровка текстов достаточно сложный процесс, который требует максимум внимания и сосредоточенности. Полная реализация проекта будет иметь важное значение как для решения фундаментальных научных задач, так и для решения актуальных прикладных задач в области сохранения фольклорного наследия калмыцкого этноса.

Литература

Азбелев С. Н. Основные понятия текстологии в применении к фольклорному материалу // Принципы текстологического изучения фольклора. М.; Л., 1966. С. 260–302.

Алтн чеежтә келмрч Боктан Шаня. Хранитель мудрости народной Шаня Боктаев / сост., предисл., коммент и прилож. Б. Б. Манджиевой; ред. Н. Г. Очирова, Б. Б. Горяева. Элиста: КИГИ РАН, 2010. 172 с. Сер.: Сокровище предков (Өвкнрин зөөр).

Буутан Санҗин туульс (Сказки Санджи Бутаева). Записи 1971–1978 годов: в 2 кн. Кн. 1. Сер.: Сокровище предков (Өвкнрин зөөр). Элиста: КИГИ РАН, 2008. 308 с.

Гацак В.М. Устная эпическая традиция во времени: историческое исследование поэтики. М.: Наука, 1989. 254 с.

Захаров В. П. Корпусная лингвистика: Учебно-метод. пособие. СПб., 2005. 48 с.

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

Лихачев Д.С. Текстология. Краткий очерк. М.; Л.: Наука, 1964. С. 69–98.

Манджиева Б. Б. Кумулятивная сказка у калмыков (синоптический анализ разновременных текстов) // Научная мысль Кавказа. 2011а. № 1(65). Ч. 2. С. 100–106.

Манджиева Б. Б. Сохранность и изменяемость типических мест в героическом эпосе «Джангар» // Монголоведение: сборник научных трудов. Вып. 5. Элиста: КИГИ РАН, 2011б. С. 287–301.

Т. С. Тягиновн амн урн үгин көрңгәс. Фольклорные материалы из репертуара Т. С. Тягиновой. Сер.: Сокровище предков (Өвкнрин зөөр). Самозапись 2004–2010 гг. / предисл. Н. Г. Очировой, сост., коммент. Б. Б. Горяевой. Элиста: КИГИ РАН, 2011. 208 с.

Убушиева Д. В. Песня «О битве богатыря Алого Хонгора с Авланги ханом» в записи от Бадмы Обушинова (к вопросам текстологии) // Вестник Калмыцкого института гуманитарных исследований РАН. 2011а. № 1. С. 168–173.

Убушиева Д. В. Текстологический анализ песен из репертуара сказителя Мукебюна Басангова// Вестник Калмыцкого института гуманитарных исследований РАН. 2011б. № 2. С. 150–153.

Хальмг улсин йөрәлмүд (Калмыцкие народные благопожелания) / сост., вступит. ст. М. Э.-Г. Эрдни-Горяева. Сер.: Сокровище предков (Өвкнрин зөөр). Элиста: КИГИ РАН, 2010. 160 с.

Чистов К. В. Фольклор. Текст. Традиция: сб. ст. / К. В. Чистов. М.: ОГИ, 2005. 272 с.

ЮНЕСКО [электронный ресурс] // URL: http://www.unesco.org/culture/languages-atlas/ (дата обращения: 30.11.2012).