МАТЕРИАЛЫ
XL МЕЖДУНАРОДНОЙ ФИЛОЛОГИЧЕСКОЙ КОНФЕРЕНЦИИ

Выпуск 24

ПОЛЕВАЯ ЛИНГВИСТИКА.
ИНТЕГРАЛЬНОЕ МОДЕЛИРОВАНИЕ
ЗВУКОВОЙ ФОРМЫ ЕСТЕСТВЕННЫХ ЯЗЫКОВ

21–25 марта 2011 г.

Санкт-Петербург

Филологический факультет
Санкт-Петербургского государственного университета

2011

ББК 81.2
М34

Ответственный редактор —
д.ф.н. А.С. Асиновский
Научный редактор —
д.ф.н. Н.В. Богданова

Материалы XL Международной филологической конференции 21–25 марта 2011 г. Вып. 24: Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков / Отв. ред. А.С. Асиновский, науч. ред. Н.В. Богданова. СПб.: Филологический факультет СПбГУ, 2011. — 240 с.

ISBN 978-5-8465-1164-4

Виктория Васильевна КУКАНОВА
(*Калмыцкий институт гуманитарных исследований РАН; Элиста*)

ОБЩАЯ СТРУКТУРА И ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ НАЦИОНАЛЬНОГО КОРПУСА КАЛМЫЦКОГО ЯЗЫКА В СВЕТЕ ПРОБЛЕМЫ РЕПРЕЗЕНТАТИВНОСТИ

Научные открытия в области вычислительной техники и ее последующее развитие способствовали появлению корпусной лингвистики как приоритетного направления современного языкознания. Первые корпусы появились практически одновременно с внедрением компьютерных технологий в гуманитарные исследования.

Структура и функционирование естественного языка всегда изучались на материале письменных или устных источников, но только в последние несколько десятилетий корпусные исследования, направленные на разработку и создание коллекций текстов на разных языках, с применением интегрированной информационной среды, оформились в самостоятельную отрасль языковедческой науки. Массовым появлением корпусных исследований на материале английского, итальянского, финского и ряда других языков (преимущественно европейских) отмечен период конца 1980-х — середины 1990-х гг. Сегодня текстовые корпусы — это мощные информационные ресурсы, которые могут быть использованы в различных исследованиях, прежде всего в лексикографии[1].

У термина *корпус* имеются два общепринятых значения, которые все же дифференцируются, в особенности в прикладной лингвистике. Различают «корпус первого порядка», под которым понимают всего лишь коллекцию текстов на некотором языке, и «лингвистический корпус» (языковой). Последнее — более объемное и системное понятие, называющее ряд признаков-критериев, которыми должна обладать та или иная коллекция текстов. Это «...большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач»[2]. Таким образом, коллекция текстов может быть названа корпусом, если она обладает ре-

презентативным объемом, представлена в электронном аннотированном виде как определенная система.

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 считаные секунды.

Поясним, что значит разметка, или аннотирование, на примере Национального корпуса русского языка — НКРЯ[3].

НКРЯ начал создаваться в 2001 г. учеными из разных городов и разных организаций. Сейчас, в 2011 г., общий объем корпуса составляет 340 млн словоупотреблений и включает в себя два массива текстов, принадлежащих разным периодам, которые условно были подразделены на современные письменные тексты и ранние тексты. Эти два массива образуют тексты разных направлений и жанров, разных функциональных стилей.

Как свидетельствует имеющаяся литература по корпусной лингвистике[4], в создании корпуса имеется немало открытых проблем, в том числе теоретического характера, в частности, проблема объема выборки, проблема репрезентативности корпуса и подкорпусов[5], проблема метаразметки[6] и т. д.

Одно из главных требований к корпусу — сбалансированность материала по конкретным характеристикам, поскольку в этом случае на «выходе» лингвист имеет более репрезентативные результаты. Доли текстов разного типа должны быть примерно одинаковыми, поскольку только в этом случае не искажается статистика.

Всем текстам должно быть дано метаописание (метаразметка), которое состоит из двух блоков:

- экстралингвистические признаки:

1) автор текста: имя, пол, дата рождения (или примерный возраст);

2) название текста;

3) время создания текста (точное или приблизительное);

4) объем текста;

- параметры текста в зависимости от рода и функционального стиля:
- художественные (прозаические и поэтические) тексты;
- нехудожественные тексты;
- драматургия.

Существуют и подкорпусы, которые создаются для решения определенных лингвистических задач. Например, синтаксический подкорпус представляет собой глубоко аннотированную систему данных по модели «Смысл ↔ Текст»[7], где показано дерево синтаксических зависимостей и классифицированы семантические роли, выраженные в словах.

Лингвистическое аннотирование — это анализ элементов текста, его составляющих, в зависимости от конкретной лингвистической задачи (анализ морфологии, синтаксиса или семантики). Каждый элемент в тексте помещен в так называемые тэги, в которых содержится характеристика слова применительно к контексту его употребления.

На современном этапе развития корпусной лингвистики существует значительное число корпусов европейских языков, однако весьма незначительно пока количество исследований и попыток создания корпусов для вымирающих или находящихся на грани исчезновения языков, в частности калмыцкого. Кроме того, следует подчеркнуть, что вопросы создания корпуса были поставлены и разрабатывались преимущественно на материале флективных языков. Поэтому привлечение данных восточных языков, в частности монгольских, как языков иного типа, представляется актуальным для расширения и углубления фактической базы при рассмотрении проблемы в общелингвистическом плане.

Принципиально новый — корпусный — подход к изучаемым явлениям неизбежно ведет к определенной корректировке или даже пересмотру ряда положений традиционной грамматики. Многие теоретические положения, выдвинутые в калмыцком языкознании, остаются до сих пор не доказанными, а предложенные решения некоторых проблем представляются весьма неоднозначными, целый ряд вопросов излагается в общем виде и нуждается в проверке на конкретном языковом материале, который, как правило, не представителен по своему объему.

Наличие основного корпуса текстов, представляющего литературный язык на определенном этапе его существования, во всем многообразии жанров и стилей, — абсолютно необходимая предпосылка для создания новой академической грамматики и академического словаря калмыцкого языка на основе его интегрального описания, предполагающего единство грамматики и лексики. Данные работы послужат базой для разработки практической, учебной грамматики и словарей разных типов, в том числе школьных, а также учебных пособий и справочников. Поэтому крайне важно приложение идей и методов корпусной лингвистики к материалу калмыцкого языка.

Такого рода проекты могут, во-первых, усилить интерес к изучению этих языков, во-вторых, облегчить работу лингвистов, уже занимающихся их исследованием, в-третьих, служить материалом для обучения языку детей. При помощи Национального корпуса калмыцкого языка (НККЯ) можно составлять различные упражнения, начиная с орфографических и пунктуационных и заканчивая упражнениями, связанными с расширением вокабулярия учащихся и его закрепления на практическом материале. Однако для достижения этой прикладной цели нужны «идеальные» тексты калмыцкого литературного языка, со снятой омонимией, орфографически и пунктуационно выверенные, не имеющие погрешностей в лексическом и граммагическом планах. Можно впоследствии также добавить модуль, отвечающий за произношение списка наиболее частотных слов, для удобства запоминания и правильного восприятия тех или иных единиц.

Калмыцкий язык, как известно, по своей структуре является агглютинативным, принадлежит к монгольской группе алтайской семьи языков. Калмыки вошли в состав Российского государства со своей письменностью «тодо бичг» ('ясное письмо'), использовавшейся вплоть до 1924 г. Затем начинается достаточно трудный этап в сфере графики: несколько раз меняли систему письма (кириллица → латиница → кириллица). Все эти события самым пагубным образом отразились на закреплении орфографических норм в калмыцком литературном языке. Калмыцкий язык представлен тремя диалектами: дербетским, легшим в основу формирования литературного языка, торгутским и бузавским,

между которыми имеются различия на фонетическом, лексическом и морфологическом уровнях.

Лингвисты Калмыцкого института гуманитарных исследований (КИГИ) РАН впервые приступили к осуществлению пилотного проекта по созданию НККЯ, а именно сбалансированной коллекции как устных, так и письменных источников, что имеет очень важное социолингвистическое значение, так как создание соответствующих ресурсов фиксирует находящийся под угрозой исчезновения миноритарный язык одного из субъектов Российской Федерации и повышает его валидность, жизнеспособность.

Сейчас происходит сбор всех доступных текстов на калмыцком языке, их сканирование и распознавание (на момент написания настоящей работы имелся «корпус первого порядка» общим объемом в 2 млн словоупотреблений).

Необходимость реализации этого фундаментального проекта очевидна, так как в Республике Калмыкия сложилась ситуация постепенной утраты калмыцкого языка, а вместе с ним — особого, специфичного, видения мира, что ведет к потере этнической идентичности личности, т. е. ее национальной принадлежности. В условиях ассимиляции калмыков русским населением, инокультурного окружения и, шире, глобализации калмыцкий язык оказывается в еще большей опасности: практически на грани исчезновения.

Переломным моментом в языковой ситуации была депортация калмыков в Сибирь (1943–1957 гг.), что и явилось началом утраты национального языка, а через него — и этнической идентичности. По данным переписи 2002 г., калмыков в России насчитывалось около 156 тыс., из них количество реальных носителей калмыцкого языка не превышает и 10 %, т. е. не более 15 тыс.[8] В настоящее время основная часть носителей калмыцкого языка — это билингвы с доминирующим русским языком. Калмыков, свободно говорящих и осуществляющих всю коммуникацию на родном языке, единицы, и проживают они исключительно в поселках, удаленных от столицы и райцентров.

В ближайшем будущем может сложиться ситуация, когда записать устную речь носителей калмыцкого языка станет уже невозможным. Что касается письменных текстов, то число авторов, пишущих на кал-

мыцком языке, неуклонно сокращается. Несмотря на многочисленные попытки государственных структур остановить процесс утраты языка, он все же продолжается.

В связи с этим актуальным становится вопрос о представительности корпуса калмыцкого языка и его сбалансированности. Оговоримся сразу, что планируемый корпус будет состоять из конечного числа текстов, так как текстов как таковых существует совсем немного и отсутствует перспектива появления новых нужных текстов. Однако для достижения основной цели корпуса — адекватного отражения лексико-грамматических фактов, типичных для всего языка и обладающих частотным характером, — достаточно, по мнению корпусных лингвистов, объема в 10–20 млн словоупотреблений[9]. Для более полного и многоаспектного описания языка необходим корпус объемом свыше 100 млн словоупотреблений. Мы отдаем себе отчет, что пока просто невозможно собрать коллекцию калмыцких текстов в таком объеме.

Итак, состав создаваемого корпуса следующий.

I. ПИСЬМЕННЫЙ КОРПУС

1. Художественные произведения, которые образуют поэтический и прозаический подкорпусы. Первый тип необходимо выделить, поскольку интонационно-ритмическая структура, создавшаяся при помощи аллитерации, а не рифмы, весьма уникальна в калмыцком языке, хотя последнее время наблюдается тенденция перехода на иную систему ритмической структуризации стихотворений, а именно на рифмование концов строк.

Особую группу здесь составляют фольклорные произведения, которые содержат большое количество архаических элементов: сказки, эпос «Джангар», исторические песни, предания и мифы, малые фольклорные жанры (пословицы и поговорки, трехстишья, загадки и др.). К тому же в подобных текстах отражается языковая структура разных диалектов калмыцкого языка, так как при расшифровке записей сохраняются все лексические и грамматические особенности языка того или иного сказителя, что является главным требованием фиксации фольклорных произведений.

2. Научные тексты составляют мизерную часть корпуса, хотя они все же присутствуют. Следовательно, существует проблема неразработанности терминологии разных научных сфер и областей, а также нет четкой дифференциации научного стиля от других функциональных стилей на грамматическом (морфологическом и синтаксическом) уровне.

3. Официально-деловые тексты представляют небольшое количество, так как официальная коммуникация в Калмыкии ведется на русском языке, что умаляет статус калмыцкого языка как государственного.

4. Публицистические тексты представлены материалами национальной газеты «Хальмг үнн», где около 50–70 % текстов — на калмыцком языке. Сотрудники редакции газеты любезно передали нам электронные текстовые файлы из своего архива за 2006–2010 гг. В настоящее время осуществляется обработка этих файлов, так как существует конфликт, связанный с кодировкой исконно калмыцких букв.

Тексты **эпистолярного подстиля**, условно выделяемого учеными как письменная разновидность разговорного стиля. Письма являются исчезающим видом коммуникации, однако пока еще их можно обнаружить в личных архивах.

В данном массиве текстов особую группу составляют параллельные тексты: это переводы Библии, художественных, в основном классических, произведений и т. п. на калмыцкий язык. В рамках создания корпуса работа будет построена на статистическом анализе разных типов и видов параллельного текста — текста на языке агглютинативного строя (калмыцком) вместе с русским переводом. Между единицами оригинального и переводного текстов (обычно — между предложениями) будет установлено соответствие с помощью процедуры выравнивания. В процессе перевода, как известно, предложения могут разделяться, сливаться, удаляться, вставляться или менять последовательность, что создает немалые трудности для идентификации предложений в битексте. Выравнивание само по себе представляется крайне сложной лингвистической задачей, имеющей, тем не менее, очень важное значение, поскольку выровненный параллельный корпус послужит основным инструментом как в дальнейших научных изысканиях[10], так и в оптимизации преподавания калмыцкого языка.

Как видно из состава корпуса письменных текстов, проблема репрезентативности и сбалансированности материала носит актуальный характер. Отсутствуют как сбалансированность текстов по функциональным стилям (превалируют тексты художественного и публицистического стилей, в то время как остальные стили представлены небольшим количеством текстов, что может привести к искаженности в некоторой степени результатов исследований), так и представительность, другими словами, объем предполагаемого корпуса оказывается весьма незначительным. С другой стороны, представленность текстов в корпусе бедна в хронологическом плане, т. е. материал в основном принадлежит современному калмыцкому языку (начиная приблизительно с 1950-х гг.). Ранние тексты, написанные на «тодо бичиг» ('ясном письме') и латинице, следует сначала перевести на кириллицу для удобства их анализа и последующей обработки, поскольку богатство языка скрыто в его истории и развитии. Отметим, что в калмыцкой лингвистике давно назрела необходимость создания исторической грамматики, которая бы дала полную картину изменения грамматических категорий.

II. УСТНЫЙ КОРПУС

Было бы интересно реализовать проект, подобный «Одному речевому дню», который был осуществлен сотрудниками Санкт-Петербургского государственного университета в 2007–2008 гг.[11] Работа по записи устных текстов должна соответствовать требованиям полевой лингвистики в условиях языкового сдвига[12]. В Лингвофольклорной лаборатории КИГИ РАН имеется фонотека записей фольклорных произведений, устной речи и записей по истории и этнографии калмыцкого народа, часть которых (приблизительно 60 %) уже оцифрована. Около 70 % всего массива записей — на калмыцком языке, остальная часть — на русском. Осуществлен пилотный проект расшифровки одной из записей[13], однако текст, полученный в результате орфографической расшифровки, был разбит на предложения, проставлены пунктуационные знаки, все признаки устной речи (протяжки, паузы хезитации, паузы, обрывы, операции отмены[14]) утеряны. Для изучения фонетических особенностей речи данные записи не подходят, так как в них присутствуют шумы разных видов, а также нарушена частота записи, но для

исследований остальных уровней языковой структуры они могут быть использованы.

Корпус должен быть наполнен текстами различной степени спонтанности. Воспользуемся классификацией, предложенной И.Н. Борисовой[15]:

1) **спонтанная**, или неподготовленная, речь: диалог и монолог;

2) **частично подготовленная**, или квазиспонтанная, речь: интервью, рассказ на заранее подготовленную тему, пересказ, описание и т. д.,

3) **подготовленная** речь: публичные выступления, чтение, пересказ, чтение стихотворений, выученных наизусть, и т. д.

Проблема представительности корпуса тесно связана с метаописанием текстов. Необходимо сбалансировать корпус и с этой точки зрения: обеспечить высокую степень представленности текстов, авторами которых являлись бы и мужчины, и женщины, люди разных возрастов и т. д. Для калмыцкого языка актуальна проблема репрезентации в корпусе разных говоров, так как имеются существенные различия между литературной и диалектной системами.

Таким образом, осуществление проекта по созданию Национального корпуса калмыцкого языка предусматривает несколько этапов:

- проведение каталогизации существующих текстов на калмыцком языке с указанием графической системы, поскольку, повторим, существуют тексты трех видов, их количество носит конечный характер, а пополнения в объемах корпуса сегодня практически нет;
- оцифровка текстов, которая решается достаточно легко в настоящее время, хотя, отметим, существуют проблемы с распознаванием источников, опубликованных в 1950–1980 гг., качество распознавания оставляет желать лучшего[16];
- доработка электронных файлов на предмет филологической выверки и корректуры текстов. Орфография и пунктуация останутся в авторском виде, хотя единообразия в написании слов и оформлении пунктуации нет;
- разметка текстов, описание дополнительных параметров (метаданные автора и самого текста, т. е. информация экстралингвистического характера).

Все эти этапы относятся к предобработке, следующие этапы должны быть связаны с автоматической обработкой текстов. Для решения проблемы автоматической обработки требуется создать лемматизатор, программы для приведения словоформы к начальной форме, для которой, в свою очередь, необходим обратный словарь калмыцкого языка, который позволит выстроить систему парадигм словоизменения[17].

Подобные корпусно-ориентированные работы успешно ведутся на материале бурятского и монгольского языков, где уже имеются определенные результаты[18]. Так, на базе корпусных данных создан частотный словарь монгольского языка Внутренней Монголии Китая[19].

Имеющиеся в нашем распоряжении электронные ресурсы пока представляют собой в лингвистическом отношении «сырой», то есть неразмеченный, текст. Поэтому основными задачами, стоящими в настоящее время перед монголоведной корпусной лингвистикой, являются, во-первых, формирование репрезентативного корпуса текстов (в том числе речевых записей), в котором были бы представлены все функциональные стили калмыцкого языка, и, во-вторых, создание в корпусе лингвистической разметки, в первую очередь морфологической и синтаксической, чтобы его можно было использовать в исследовании текстового поведения тех или иных языковых явлений.

Реализация этого фундаментального проекта откроет перспективы дальнейшей, более детальной, разработки вопросов, связанных с исследованием калмыцкого языка. Это не только фундамент для лингвистических исследований будущего, но и один из способов сохранения этнической культуры, ибо язык есть часть культуры народа, ее базис и синкретичное выражение.

Естественно, осуществление столь грандиозного проекта по созданию НККЯ невозможно без государственной поддержки в виде грантов, дополнительного финансирования, юридических консультаций и др., которая должна осуществляться на всех уровнях, начиная с республиканского и заканчивая международным.

Примечания

[1] См. подробно: Крылов С.А. Стратегии применения интегрированной информационной среды StarLing в корпусной лингвистике и в компьютерной лексикографии // Orientalia et classica. Труды Института восточных культур и античности. Вып. XIX. Аспекты компаративистики / Смирнов И.С. (ред.). М., 2008. С. 649–668.

[2] Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. СПб., 2005. С. 4.

[3] Национальный корпус русского языка [Электронный ресурс] // URL: http://ruscorpora.ru/ (15.04.2011); см. также: 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., 2005. С. 193–214.

[4] Čermák F. Today's Corpus Linguistics: Some Open Questions // International Journal of Corpus Linguistics. Vol. 7. № 2. 2002. P. 265–282; Вербицкая Л.А., Казанский Н.Н., Касевич В.Б. Некоторые проблемы создания национального корпуса русского языка // Научно-техническая информация. Сер. 2. 2003. № 6. С. 2–8.

[5] Шаров С.А. Представительный корпус русского языка в контексте мирового опыта // Научно-техническая информация. Сер. 2. 2003. № 6. С. 9–17; Шаров С.А., Савчук С.О. Типология текстов для представительного корпуса // Труды международной конференции «Корпусная лингвистика – 2004» (Санкт-Петербург, 11–14 октября 2004 г.). СПб., 2004. С. 352–362.

[6] Волков С.С., Захаров В.П., Дмитриева Е.А. Матеразметка в историческом корпусе XIX века // Труды международной конференции «Корпусная лингвистика – 2004» (Санкт-Петербург, 11–14 октября 2004 г.). СПб., 2004. С. 86–98; Савчук С.О. Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005. С. 62–88.

[7] Мельчук И.А. Русский язык в модели «Смысл↔текст». М.; Вена, 1995.

[8] Итоги Всероссийской переписи населения 2002 года по Республике Калмыкия. Элиста, 2004. С. 49.

[9] Корпусная лингвистика [Электронный ресурс] // URL:http://ru.wikipedia.org/wiki/Корпусная лингвистика (25.04.2011).

[10] Например: исследование грамматических и лексических трансформаций, лексических соответствий и различий в переводной структуре предложений, а также для теории перевода разноструктурных языков и индивидуальных особенностей переводных стратегий, мотивации выбора той или иной структурной единицы.

[11] Архипова Е.А., Богданова Н.В., Бродт И.С., Маркасова Е.В., Куканова В.В., Шерстинова Т.Ю. О корпусе звучащих текстов на русском языке (формирование учебных материалов нового типа) // Текст: проблемы и перспективы. Аспекты изучения в целях преподавания русского языка как иностранного. Материалы IV Международной научно-практической конференции (22–24 ноября 2007 г.). М., 2007. С. 31–34; Богданова Н.В., Бродт И.С., Куканова В.В., Павлова О.В., Сапунова Е.М., Филиппова Н.С. О «корпусе» текстов живой речи: принципы формирования и возможности описания // Компьютерная лингвистика и интеллектуальные технологии. Вып. 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А.Е. Кибрик. М., 2008. С. 57–61 (http://www.dialog-21.ru/dialog2008/materials.asp?type=reports (15.04.2011)).

[12] Кибрик А.Е. Методика полевых исследований (к постановке проблемы). М., 1972.

[13] Буутан Санжин туульс (Сказки Санджи Бутаева). Записи 1971–1978 гг. В 2 кн. Кн. 1. Сер.: Өвкнрин зөөр. Элиста, 2008.

[14] Филлипова Н.С. Операция отмены как способ организации спонтанной речи (на материале устных спонтанных монологов-описаний) // Материалы XXXVI международной филологической конференции. Вып. 20. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 12–17 марта 2007 года / Отв. ред. А.С. Асиновский, Н.В. Богданова. СПб., 2007. С. 86–91; см. также: Она же. Принципы построения устного описательного дискурса (на материале русской спонтанной речи). Дис. ... канд. филол. наук. СПб., 2010 (машинопись).

[15] Борисова И.Н. Русский разговорный диалог: структура и динамика. Екатеринбург, 2005. С. 132–144.

[16] В ходе эксперимента было выявлено, что тексты имеют достаточное количество ошибок и погрешностей, их исправление задерживает процесс оцифровки.

[17] Куканова В.В. К постановке проблемы создания обратного словаря калмыцкого языка (предварительные замечания) // Монголоведение: сборник научных трудов. Вып. 5. Элиста, 2011. С. 199–206.

[18] См., например, работы: Бадагаров Ж.Б. О репрезентативности текстов и элементах программного инструментария для корпуса бурятского языка // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам. El' Manuscript-08. Материалы Международной научной конференции (Казань, 26–30 августа 2008 г.). Казань, 2008. С. 28–31; Бадмаева А.Д., Бадагаров Ж.Б., Цыдыпов Б.З. Общие проблемы формирования корпуса бурятского языка // Труды Международной конференции «Корпусная лингвистика – 2008». СПб, 2008. С. 24–30; Бадмаева А.Д. Бурятский язык и корпусная лингвистика // Состояние и перспективы развития бурятского языка. Материалы форума бурятского языка. Улан-Удэ, 2009. С. 83–86; Она же. Корпус бурятского языка: проект [Электронный ресурс] // http://www.globecsi.ru/Articles/2008/Badmaeva3.pdf (20.04.2011); Ринчинов О.С. Корпус бурятского языка и прикладные задачи компьютерной лингвистики // Состояние и перспективы развития бурятского языка. Материалы форума бурятского языка. Улан-Удэ, 2009. С. 88–89; Бадмаева А.Д., Badam-Osor Kh., Fujii A. A Lemmatization Method for Modern Mongolian and its Application to Information Retrieval [Электронный ресурс] // http://www.if-lab.slis.tsukuba.ac.jp/ fujii/paper/ijcnlpkhab.pdf (25.04.2011).

[19] Bagatur Da., Djirumt Bu. Oduüy-e-yin mongul kelen-ü üge-yin dabtamji-yin toli. Kök xot: Ober mongul-un surgan kümüjil-ün keblel-ün küriy-e, 1998.