О репрезентативности текстов и элементах программного инструментария для корпуса бурятского языка[1]

Л. Д. Бадмаева, Ж. Б. Бадагаров

Институт монголоведения, буддологии и тибетологии СО РАН, Бурятский государственный университет, Улан-Удэ, Россия

Технологии корпусной лингвистики уже широко признаются наиболее оптимальными в исследовательской практике современной лингвистики. Любой язык, будь он живой или сохранившийся только в письменных памятниках, сегодня может быть подвергнут «корпоризации», что безоговорочно предоставляет заманчивые возможности исследователям самых разных отраслей знания. Поэтому и такой язык, как бурятский, также может быть включен в орбиту корпусных технологий с целью создания условий для его дальнейшего функционирования в электронных системах как в исследовательской, так и образовательной сферах.

Методология развивающейся корпусной лингвистики позволяет получать максимально надежные языковые факты реального функционирования языка. В основном посредством сканирования[2], а также на основе договоров с издательствами создается архив текстовых материалов для корпуса бурятских текстов, на котором апробируются конкордансеры [Grieves; Watt], позволяющие получать полноконтекстную информацию о языковых фактах. Материалы поиска дают возможность формировать разного рода классы рассматриваемых языковых единиц.

Традиционные подходы не позволяют просматривать и анализировать все тексты в достаточно полной мере. Даже при выполнении длительных работ по накоплению эмпирического материала список текстовых источников в монографических работах (книгах, диссертациях, статьях) страдает определенной ограниченностью (данная ограниченность ни в коем случае не является упреком в адрес исследователей, она объясняется тем, что в период подготовки их научных работ компьютерные технологии не были

[1] Работа выполнена при финансовой поддержке РФФИ (проект 08-06-00151-а) и Американского совета научных сообществ (ACLS).

[2] При малочисленности бурятоязычных сайтов в Интернете сканирование представляется основным средством преобразования бумажных версий текстов в электронные.

развиты так, как в настоящее время) как в количественном плане, так и в качественном. Под последним мы подразумеваем репрезентативность самого списка источников, который имеется в исследованиях по бурятскому языкознанию. Например, можно взять списки источников, указанные в трудах наиболее крупных исследователей по бурятскому и шире – монгольскому языкознанию, например, в работах Ц. Б. Цыдендамбаева, А. А. Дарбеевой, Л. Д. Шагдарова, Г. Ц. Пюрбеева, В. И. Рассадина, И. Д. Бураева, Ц. Ц. Цыдыпова, В. И. Золхоева, С. Л. Чарекова, Д. Д. Доржиева, Э. Р. Раднаева, У.-Ж. Ш. Дондукова, Ц. Б. Будаева, М. Н. Орловской, С. Ш. Чагдурова, Д. А. Алексеева, Д. Д. Амоголонова, Ш-Н. Р. Цыденжапова, Л. К. Скрибник и др. Мы ориентируемся на работы в основном советского (начиная со второй половины XX века) и постсоветского периодов как являющиеся наиболее определяющими становление и поступательное развитие монголоведения в целом и бурятоведения в частности.

При всей фундаментальности трудов данных авторов рассматриваемый список использованных источников для эмпирического материала будет иметь определенные границы. Данные границы будут либо временнымми, либо стилистическими, либо диалектологическими, либо иными. Ясно, что данные списки не будут полностью совпадать. Некоторые источники будут встречаться в большинстве перечней («кочевать» из одного исследования в другое), некоторые будут встречаться только в каком-то одном списке. Наличие во всех перечнях какого-то источника может стать критерием обязательного его включения в корпус. Интересны также те тексты, которые имеют единичное употребление только в каком-либо одном списке. Если объединить все перечни текстов, указанных в данных исследованиях, то, вероятно, текстовые материалы такого обобщенного списка можно принять за источники, которые в первую очередь необходимо включить в бурятский корпус. В данных списках в целом за счет единичности «покрываются» в той или иной степени все поля функционирования исследуемого в них бурятского языка. Естественно, при создании корпуса текстов на бурятском языке нельзя ограничиваться только данным обобщенным списком текстов, поскольку в настоящее время постоянно порождаются новые тексты, что связано еще и с

тем, что не быстро, но регулярно разрабатываются бурятоязычные сайты, на которых в будущем будут появляться электронные версии текстов на бурятском языке без промежуточных их публикаций на бумажных носителях. Исходя из этого, бумажные версии в настоящее время можно назвать промежуточным этапом функционирования текстов на бурятском языке, т. к. они постепенно приобретают новую жизнь в электронном виде. Надо отметить, что электронные бурятские тексты стали появляться и на компакт-дисках, правда, такой тип публикации пока не очень распространен и популярен. Таким образом, бумажные версии бурятских текстов приобретают новую, до сих пор неизвестную для них, электронную жизнь. И она открывает новые горизонты для научного поиска не только лингвистам, но и исследователям других специальностей.

Подготовленный нами электронный архив бурятских текстов является в основном полнотекстовым, что обусловлено территориальным ограничением функционирования самого бурятского языка. При создании архива текстовых материалов учитывается принцип репрезентативности корпуса, который должен отражать основные стили современного бурятского литературного языка – художественно-литературный, публицистический и учебно-научный, с преобладанием текстов, реализующих первый стиль. Репрезентативность корпуса основывается на представленности современного бурятского языка разными по жанру и стилю текстами, опубликованными со второй половины XX века по настоящее время. Сейчас работы по комплектации текстовых материалов для формирования электронного корпуса выполнены в объеме одного миллиона лексических вхождений - словоформ. Технологии корпусной лингвистики в значительной степени дополняют методы и приемы традиционной лингвистики, никоим образом не умаляя значимости традиционных исследований и способствуя расширению исследовательских возможностей, что в свою очередь, без сомнения, в дальнейшем поднимет бурятоведение на качественно новый уровень. Результатом подобного дополнения должны стать конкретные лингвистические описания, построенные на фактическом материале, анализ функционирования которого носит глобальный характер по отношению к репрезентативному корпусу.

Технологии корпусной лингвистики направлены на своего рода «глобализацию» исследования в пределах того или иного корпуса, включающего оригинальные тексты на данном языке. Анализ материала, рассматриваемого глобально в определенном корпусе, способствует максимальному выявлению его характерологических черт, что позволяет осуществлять наиболее достоверное его описание.

Основными направлениями работы для организации корпуса бурятского языка в настоящее время являются следующие: экстралингвистическая разметка, графематический анализ, морфологическая разметка [3] . Для добавления экстралингвистической информации (метаданные: формальное описание текста — автор, название, время и место создания, объем текста; содержательное описание текста — жанр, тип, хронотоп текста) будет использована программа UAM CorpusTool. При решении задачи графематического анализа предполагается разработка соответствующего программного инструментария, учитывающего особенности бурятских текстов. В частности, одной из таких особенностей является широкое употребление в бурятском языке парного основосложения как одного из словообразовательных средств. Для морфологической разметки (POS-tagging) бурятских текстов требуется разработка программ для автоматической лемматизации и разметки. Следует отметить, что исследования в этом направлении только начинаются. Одним из предварительных результатов является создание программы, выполняющей автоматическую генерацию падежных форм имен бурятского языка. В дальнейшем планируется создание словаря основ и словаря аффиксов для создания парсера именных словоформ. Определен инвентарь грамматических помет (граммем), который предназначен для использования в корпусе и который будет уточняться по мере необходимости. В целом, говоря о стандарте разметки, можно сказать, что нами будут учтены требования международного стандарта представления электронных текстов TEI. Для создания собственного подмножества TEI используется web-программа TEI Pizza Chef.

[3] Задача синтаксической и семантической разметки пока отложена на будущее.

Ввиду того, что программы-конкордансеры ограничены в своих возможностях, для более полного использования данных корпуса будет применен корпусный менеджер. Выбор будет сделан из существующего на данный момент разнообразия готовых решений, таких, как Manatee / Bonito, IMS Corpus Workbench (CWB), CQP/CQP-WS, Xaira (SARA).

Список литературы

Grieves – *Grieves, Chris*. ConcApp Concordancing Programs [Электронный ресурс]. – Режим доступа : www.edict.com.hk./PUB/concapp, свободный. – Загл. с титул. страницы.

Watt – *Watt, Robert*. Concordance [Электронный ресурс]. – Режим доступа : www.concordancesoftware.co.uk, свободный. – Загл. с титул. страницы.

On the representativity of texts and software development for a corpus of the Buryat language

Lubov' D. Badmaeva, Zhargal B. Badagarov

Institute for Mongolian, Buddhist, and Tibetan Studies of the Siberian Branch of the Russian Academy of Sciences, Buryat State University, Ulan-Ude, Russia

A corpus of Buryat texts is being constructed, mostly by scanning. Concordance software is used on these electronic texts, which include samples of literary, public, and scholarly works. Extralinguistic analysis, morphological annotation, and graphematic analysis are discussed.