УДК 811.512.31, 81’33
DOI 10.18101/978-5-9793-1709-0-227-229

О РАСШИРЕНИИ КОРПУСА БУРЯТСКОГО ЯЗЫКА

© **Бадмаева Любовь Дашинимаевна**
кандидат филологических наук, доцент, ведущий научный сотрудник,
Институт монголоведения, буддологии и тибетологии СО РАН
Россия, г. Улан-Удэ
ldbadm@gmail.com

В статье описаны работы по обновлению корпуса бурятского языка в 2021 г. Важнейшими результатами проделанных работ являются: база данных текстов объемом более 400 тыс. словоупотреблений; увеличение жанрового разнообразия корпуса за счет фольклорных текстов; обновленный корпус бурятского языка на усовершенствованной платформе Цакорпус объемом 2,8 млн словоупотреблений; дополнительные интерфейсы на бурятском и английском языках; увеличение морфологической разметки словоформ корпуса до 76%.
Ключевые слова: бурятский язык, корпусная лингвистика, база данных, текст, метаразметка, морфологическая разметка.

Работа выполнена в рамках государственного задания (проект «Мир человека в монгольских языках: анализ средств выражения эмотивности») № 121031000258-9 и при финансовой поддержке Минобрнауки Республики Бурятия (контракт №121-2021, договор №125-2021).

Благодарность. Сотрудники ИМБТ СО РАН Л. Д. Бадмаева, Ю. Д. Абаева, Г. Н. Чимитдоржиева, О. С. Ринчинов, выполнившие работы по контракту и договору, сердечно благодарны Т. А. Архангельскому (Университет Гамбурга, Германия) за постоянную поддержку на высоком профессиональном уровне при обновлении и функционировании бурятского корпуса; С. А. Крылову (Институт востоковедения РАН, Москва) за ценные консультации при подготовке материалов.

ON EXPANSION OF THE BURYAT LANGUAGE CORPUS

Liubov D. Badmaeva
Cand. Sci. (Phil.), A/Prof., Leading Researcher,
Institute for Mongolian, Buddhist and Tibetan Studies
of the Siberian Branch of the Russian Academy of Sciences
Ulan-Ude, Russia
ldbadm@gmail.com

The article deals with the description of works on updating the Buryat language corpus. The corpus was updated to 2.8 million word usages. The morphological word form tagging of the corpus has been increased to 76%.
Keywords: the Buryat language, corpus linguistics, database, text annotation, morphological tagging.

The work was carried out within the state order (project "The World of Man in the Mongolian Languages: Analysis of the Means of Expressing Emotivity") No. 121031000258-9 and with the financial support of the Ministry of Education and Science of the Republic of Buryatia (contract No. 121-2021, contract No. 125-2021).

Gratitude. Employees of the IMBT SB RAS L. D. Badmaeva, Yu. D. Abaeva, G. N. Chimitdorzhieva, O. S. Rinchinov, who performed the work under the contract and the agreement, are sincerely grateful to T. A. Arkhangelsky (University of Hamburg, Germany) for constant support at a high professional level in the renovation and functioning of the Buryat corpus; S. A. Krylov (Institute for Oriental Studies, Russian Academy of Sciences, Moscow) for valuable consultations in material preparation.

Цель работы (1) заключалась в расширении и доработке справочно-информационного онлайн ресурса по литературному бурятскому языку, который (назовем его — старый корпус) остался доступным по прежнему адресу [Бурятский корпус]. Для достижения вышеназванной цели выполнялись такие задачи, как подготовка новой текстовой базы данных со всеми сопутствующими процедурами, которые необходимы при дигитализации текстов, и их метаразметка.

В составляемом нами (2) ресурсе основные цели и задачи направлены на тексты, опубликованные в издательствах, отвечающие нормам издательской деятельности и отражающие современный бурятский язык как «объективную речевую действительность» [Богоявленская, 2016, с. 163] в соответствии с его общепринятой академической грамматикой.

Результатом подборки художественных текстов, отсутствовавших в старой версии Корпуса, являются разные художественные и фольклорные жанры: 4 романа, более 30 рассказов, 1 повесть, 2 очерка, 22 сказки. Увеличение жанрового разнообразия текстовой базы данных Корпуса способствует повышению репрезентативности и сбалансированности содержательной стороны Корпуса. В числе названных материалов 23 рассказа и очерки принадлежат перу классика бурятской литературы Х. Намсараева, написанные в период с 1928 г. до 1955 г. Язык произведений классика отражает богатство бурятского языка, что бесспорно признается бурятоведами, несмотря на обусловленность тематики, содержания его произведений беспрекословными догмами советской идеологии того периода.

Совершенно справедливо то, что «сбор текстов является наиболее времязатратным шагом при создании крупных корпусов. Составление репрезентативных корпусов требует включения в корпус разных жанров художественной литературы и публицистики разных временных периодов, а это, в свою очередь, требует гигантской работы по сканированию, оптическому распознаванию и вычитке текстов» [Архангельский, 2019, с. 530].

Для того чтобы текстовые материалы стали полезными для различных пользователей, в первую очередь для представителей научной и образовательной сфер (филологи различных специальностей и категорий — лингвисты, литературоведы, фольклористы, этнографы, историки, учащиеся, студенты), далее для тех, кто интересуется бурятским языком с сугубо практическими потребностями его использования как в профессиональной деятельности, так и в повседневной жизни (журналисты, писатели, обычные граждане), требуется проведение специальных работ по конвертации текстов по различным параметрам в компьютерных программах.

С подготовленными текстами проведены такие процедуры, как автоматическая обработка текстов и морфологического анализатора. Итогом комбинированной обработки подготовленных по контракту текстовых данных и их интеграции в базу данных старого корпуса является загрузка на новом сайте дополненного и обновленного Корпуса [Корпус бурятского языка].

Важнейшими результатами проделанных работ являются: новая база данных текстов объемом более 400 тыс. словоупотреблений, включающая оригинальные бурятские и переводные произведения. Впервые в Корпус вводятся фольклорные тексты в виде сказок, являющиеся отражением богатой народной речи. Для обогащения жанрового состава Корпуса впервые включены сказки, перевод исторического романа А. Бальбурова «Поющие стрелы» (*Зэдэлээтэ зэбэнүүд*) с русского языка на бурятский, выполненный литературными переводчиками. Начало формирования бурятской литературы на литературном языке относится к середине 1-й половины XX в. До настоящего времени общий период создания бурятских литературных произведений можно обозначить приблизительно в 80 лет. За такой небольшой временной период численность произведений на бурятском языке со времени начала использования кириллической графики нельзя сравнить, например, с русской литературой, у которой начало уходит корнями в прошлые века. В связи с этим с уверенностью можно сказать, что при включении в разрабатываемый корпус наряду с

оригинальными текстами литературных переводов на современный бурятский с других языков вероятным будет объем приблизительно в 10 млн словоупотреблений. Данный объем лингвистического корпуса для языков национальных республик, таких как бурятский, рассматривается корпусными лингвистами как оптимальный, хотя это не значит, что указанным показателем следует ограничиваться. Для сравнения можно привести пример с Национальным корпусом русского языка, который насчитывает уже более 1 млрд слов [НКРЯ].

Итогом выполненной работы по контракту является расширенная и дополненная версия Корпуса на новой, усовершенствованной платформе Цакорпус (разработчик — Т. А. Архангельский) объемом 2,8 млн словоупотреблений. Дата обновления Корпуса — 14.12.2021. Также важнейшим результатом является то, что Корпус получает в опытном режиме дополнительные интерфейсы на бурятском и английском языках. В силу языковой специфики (отсутствия в бурятском языке компьютерных и иных терминов в полном объеме), некоторые из них остаются непереведенными. Для современного бурятского языка характерны длинные, описательные, многословные переводы отдельных международных терминов. Для подобных переводов терминов на бурятский язык, иначе — длинных словосочетаний для выражения компьютерной команды в табличных поисковых формах, недостаточно места на сайте. Данное обстоятельство стало технической причиной использования для бурятского интерфейса переводов отдельных команд с русского языка в сокращенном варианте. По мере апробации новой версии Корпуса будут проводиться работы по отладке его системного устройства. Дополнение нового бурятского интерфейса предоставляет удобство его использования, кто недостаточно или совсем не владеет русским языком, например, буряты и монголы Монголии и Китая при их знании кириллической графики соответственно.

Обновление баз данных Корпуса служит расширению исследовательской тематики в бурятском языкознании на материале корпусных данных. Корпус дает огромное преимущество филологам-исследователям при сборе языковых фактов для анализа при подготовке словарей, научных статей, монографических работ, бакалаврских и магистерских выпускных квалификационных работ, диссертаций.

Корпус доступен пользователям для широкого использования без ограничений в свободном режиме, с необходимостью доступа к Интернету. Это дает возможность, например, использовать его в преподавательской работе на занятиях, лекциях, уроках, во время подготовки домашних заданий студентами, учащимися. Расширены возможности использования Корпуса в преподавании родного языка учащимся как старшего, так и младшего возраста, при изучении бурятского языка.

Корпус языка, как правило, — это долгосрочный проект, который необходимо периодически совершенствовать в соответствии с уровнем стремительно развивающихся технологий компьютерной и корпусной лингвистики.

Литература

1. Архангельский Т. А. Интернет-корпуса финно-угорских языков России // Ежегодник финно-угорских исследований / ФГБОУ ВО »Удмуртский государственный университет». 2019. Т. 13, № 3. С. 528–537.

2. Богоявленская Ю. В. Репрезентативность лингвистического корпуса: метод верификации достоверности полученных данных // Политическая лингвистика. 2016. № 4 (58). С. 163–166.

3. Бурятский корпус. URL: // http://web-corpora.net/BuryatCorpus/search/?interface_language=ru (дата обращения: 02.02.2022).

4. Корпус бурятского языка. URL//: http://buryat.web-corpora.net/index.html (дата обращения: 02.02.2022).

5. НКРЯ — Национальный корпус русского языка. URL: // https://ruscorpora.ru/new/ (дата обращения: 02.02.2022).