РУССКИЙ ЯЗЫК. ЯЗЫКИ НАРОДОВ РОССИИ / RUSSIAN LANGUAGE. LANGUAGES OF THE PEOPLES OF RUSSIA

DOI: https://doi.org/10.60797/RULB.2024.55.11

КОРПУСНЫЕ РЕСУРСЫ БУРЯТСКОГО ЯЗЫКА: СОСТОЯНИЕ, ПРОБЛЕМЫ, ПЕРСПЕКТИВЫ

Научная статья

Бадмаева Л.Д.[1, *]
[1] ORCID : 0000-0003-3326-9869;
[1] Институт монголоведения, буддологии и тибетологии СО РАН, Улан-Удэ, Российская Федерация

* Корреспондирующий автор (ldbadm[at]gmail.com)

Аннотация

Данная работа посвящена рассмотрению предыстории, истории организации и составления корпусных ресурсов на материале бурятского языка. Представлена характеристика задела для развития корпусного направления в бурятском языкознании, как конкордансы, частотные словари, картотеки, составлявшиеся в докомпьютерную эпоху. Дается описание состояния основного «Бурятского корпуса», Параллельного бурятско-русского корпуса и Диахронического корпуса бурятского языка, характеризуются их структура и основные проблемы при их разработках. Автором делаются выводы о высокой востребованности разрабатываемых корпусов и значении расширения в них разных видов разметок для проведения продуктивных лингвистических и прикладных исследований. Представлены перспективы рассматриваемых корпусов и основные задачи их дальнейшего развития.

Ключевые слова: бурятский язык, корпус, разметка, грамматический словарь.

CORPUS RESOURCES OF THE BURYAT LANGUAGE: STATE, PROBLEMS, PROSPECTS

Research article

Badmaeva L.D.[1, *]
[1] ORCID : 0000-0003-3326-9869;
[1] Institute for Mongolian, Buddhist and Tibetan Sciences, Siberian Branch, Russian Academy of Sciences, Ulan-Ude, Russian Federation

* Corresponding author (ldbadm[at]gmail.com)

Abstract

This work is dedicated to the prehistory, history of organization and compilation of corpus resources in the Buryat language. The paper characterizes the development of corpus resources in Buryat linguistics, such as concordances, frequency dictionaries, and cartographies compiled in the pre-computer era. The author describes the state of the main "Buryat Corpus", the Parallel Buryat-Russian Corpus and the Diachronic Corpus of the Buryat language, characterizes their structure and the main problems in their development. The author draws conclusions about the high demand for the corpuses under development and the importance of expanding different types of markups in them for productive linguistic and applied research. The prospects of the corpuses under consideration and the main tasks of their further development are presented.

Keywords: Buryat language, corpus, markup, grammatical dictionary.

Введение

В настоящее время корпусная лингвистика без сомнения является одним из актуальных направлений современного языкознания. Она прочно заняла, благодаря своим теоретическим и различного рода прикладным, инжиниринговым результатам, лидирующие позиции, как неуклонно развивающийся раздел филологической науки в целом на междисциплинарных стыках. Вышесказанное можно подтвердить и вниманием, оказывающимся со стороны государственных структур, как например:

1. в начале апреля 2024 г. на заседании Президиума Российской академии наук обсуждались задачи корпусных исследований языков и заслушивались ряд докладов по Национальному корпусу русского языка, а также по другим языкам народов России [1]; 2. в середине мая 2024 г. Домом народов России при поддержке Федерального агентства по делам национальностей России проведена I-я стратегическая сессия «Информационные технологии и языки народов России» [2], на которой рассматривались вопросы государственной национальной политики в сфере поддержки языков народов России в киберпространстве.

Методы и принципы исследования

Идеи, методы, методика корпусной лингвистики начали проникать в бурятское языкознание и развиваться, использоваться в нем примерно в то же время, как это происходило и в развитии других частных лингвистических направлений по малым и средним языкам народов России. В бурятском языкознании инструментарии корпусной лингвистики, используемые сегодня, как само собой разумеющееся, например, такие, как конкордансы, частотные словари (ЧС) привлекли внимание уже в 80-е годы прошлого века [3]. В 1992 г. до появления современных компьютеров был составлен и опубликован Г.А. Дырхеевой первый частотный словарь бурятского языка на материале текстов произведений Х. Намсараева [4], составленный на ЭВМ того периода. Исходя из сказанного, можно видеть, что бурятоязычные тексты для лингвистических целей впервые были подвергнуты машинной обработке в конце 80-х и в 90-е годы XX в. Тем не менее ввиду того, что работы выполнялись в докомпьютерную эпоху, большие их объемы

обрабатывались вручную с огромными трудовыми и временными затратами, доходящими до нескольких лет. Автор ЧС бурятского языка в 1992 г. уже тогда справедливо указывает, что практическое его использование исключительно важно для автоматической обработки текстов (АОТ) и их информации (подчеркнуто нами – Л. Бадмаева) [4, С. 7]. В тот период в отделе языкознания был задел в виде картотеки бурятского языка, занимающий не меньше десятка крупных каталожных шкафов, стоящих в кабинете вдоль двух противоположных стен длиной по 6 м. Объем данной картотеки составлял (она сохранена в фонде ЦВРК ИМБТ СО РАН) чуть более 1 миллиона карточек со словарными статьями бурятского языка, в которых записаны словоупотребления из художественной литературы с контекстами. Это – традиционная словарная картотека докорпусной эпохи. Сама картотека составлялась на протяжении десятков лет разными поколениями языковедов отдела и по прошествии времени ее использование стало гигиенически трудным в силу накопившейся многолетней пыли, а к настоящему периоду элементарно устарело.

С появлением и активным проникновением технологий корпусной лингвистики, как известно, являющейся частью компьютерной, в российскую лингвистическую среду и вслед за появлением в открытом доступе Национального корпуса русского языка (НКРЯ) с 2002 г. языковеды по другим языкам народов России стали также постепенно друг за другом ставить и решать задачи разработок своих языковых корпусов, явившихся информационными системами соответственно абсолютно нового поколения.

В разработке «Бурятского корпуса», можно сказать, результаты вероятностно-статистических методов и подходов и сами данные методы с приходом корпусных технологий приобрели свое закономерное и эффективное воплощение.

Основные результаты

Следует сказать, что разработка «Бурятского корпуса» претерпела несколько версий, которые были представлены в онлайн. С 2011 г. была открыта первая опытная версия с названием «Корпус бурятского языка» объемом около 800 тыс. словоупотреблений по адресу ЦВРК ИМБТ СО РАН [5], на котором были размещены конкордансы к бурятским художественным текстам с указанием их авторства и названий (разработчик сайта и программы конкорданса – О.С. Ринчинов). Данная версия была стабильно доступна для пользователей / исследователей первые три года, в настоящее время – эпизодически.

Опыт работ по составлению названной выше версии корпуса вместе с языковыми и программными материалами послужил нашему участию на конкурсной основе в Программе Президиума РАН запущенной в 2011 г. по фундаментальным научным исследованиям под названием «Корпусная лингвистика».

Далее, при поддержке данной Программы в 2012 г. был впервые открыт полноправный корпусный сайт по бурятскому языку [6] наряду с другими, по сути, первыми корпусами по нескольким языкам народов России. Корпусный менеджер, называющийся также и платформой, на котором представлены корпусы по Программе РАН заимствован по согласованию от разработчиков компании Corpus Technologies и Восточно-армянского национального корпуса (ВАНК). Данная платформа, представляет собой значительно усложненную по сравнению с конкордансером программу, как управления, так и использования того или иного языкового корпуса с возможностью его совершенствования и развития. К настоящему времени подобных платформ имеется уже множество. На корпусной платформе «Бурятского корпуса» имеются свой интерфейс, инструментарии, как, поисковая строка с выбором по словоформе, по лемме и некоторым грамматическим характеристикам, при этом, с возможностью выборки конкретных текстовых материалов, с настройками представления результатов поиска, выбора определенных текстовых документов, включенных в корпус. Также есть виртуальная клавиатура с тремя парами сугубо бурятских символов / букв.

В дальнейшем, данная версия нашего корпуса претерпела обновление и пополнение в 2016 и в 2021 гг., оба раза - при финансовой поддержке по Контрактам Минобразования и науки Республики Бурятия. В «Бурятском корпусе» интегрированы соответственно базы данных текстов, грамматического словаря, а также, основная часть бурятско-русского словаря. По текстовой базе данных есть возможность осуществлять поиск лексем, как по всем входящим текстовым документам, так и при необходимости – отдельным текстам. Грамматический словарь (далее – ГС) первоначально составлялся нами вручную в Excel на базе сформированного электронного словника бурятско-русского словаря [7], [8] затем на базе частотного словаря первоначальной версии «Бурятского корпуса» (составитель частотного словаря – О.С. Ринчинов). В грамматическом словаре вручную выполняется морфологическое описание со словоизменительными парадигмами. Данное описание используется в системе автоматического морфологического анализатора UniParser (разработчик – Т.А. Архангельский) для разметки / аннотации словоформ корпуса. В результате в корпусе на выданных при поиске языковых данных при необходимости посредством наведения курсора на конкретное слово всплывает мини-окошко с грамматическими характеристиками. Данные бурятско-русского словаря активируются при наборе русского слова в поисковой строке (затем, кликается кнопка «перевод» и нажимается – «искать». При наличии данного русского слова в бурятско-русском словаре корпуса, будет выдан результат с бурятской лексической параллелью в контексте своего использования. Кроме морфологической разметки в «Бурятском корпусе» есть метаразметка с указанием внешних данных текстовых документов – название, авторство, год издания, что можно видеть при каждом языковом примере употребления. При разметке внешних данных текстов использовалась также СУБД StarLing при поддержке С.А. Крылова (ИВ РАН).

В результате обновления и пополнения бурятского корпуса, выполненных к декабрю 2021 г., была открыта версия на другой корпусной платформе, усовершенствованной по отношению к предыдущей под названием Цакорпус [9] (можно отметить, что предыдущая, вышеописанная версия нашего корпуса осталась доступной по прежнему адресу). Объем корпуса увеличился на 400 тыс. словоупотреблений, в результате, общий их объем стал 2,8 млн словоупотреблений. «Бурятский корпус» получил новый интерфейс, при этом на трех языках: русском, бурятском и английском. Помимо того, что (подчеркнуто нами – Л.Б.) есть в предыдущей, то есть старой версии, на новой платформе Цакорпус в поиске появилась возможность найти несколько слов внутри одного предложения. Здесь можем подтвердить справедливые слова автора платформы Цакорпус о полезности вышеописанной возможности при

изучении сочетаемости слов и грамматических конструкций [9, С. 23]. Количество контекстов на странице выдачи выросло до 100, на прежней платформе – 50. Доля морфологического разбора составляла 76%, а к данному времени автоматическим разбором доведена до 80%. Оставшиеся проценты морфологически неразобранных слов размечаются нами теперь в онлайн в специальной программе для пополнения грамматического словаря (автор – Т.А. Архангельский). Кроме грамматических признаков к словам, приписываются в данной программе и значения, пополняя таким образом встроенный в разрабатываемом корпусе бурятско-русский словарь. В программе для пополнения грамматического словаря предусмотрена проверка введенных данных, после которой, они закрепляются в нем и происходит обновление с пополнением морфологического разбора и двуязычного словаря в масштабах корпуса. Морфологическая разметка является основой для других видов разметки – словообразовательной, синтаксической, семантической и т.д.

Еще одним современным программным инструментарием для изучения бурятского языка является «Параллельный бурятско-русский корпус» [10] объемом в 400 тыс. словоупотреблений с метатекстовой разметкой. В параллельном корпусе бурятского языка представлены выровненные тексты оригинальных бурятских произведений с литературными переводами на русский язык, а также есть обратное направление перевода – оригинальный русский текст (Пушкин А.С. Капитанская дочка) с переводом на бурятский. Развитие данного вида корпуса повысит его востребованность, например, в плане исследования проблем переводоведения, а также в деле составления билингвальных словарей различного типа. Одновременно – параллельный корпус имеет важное значение для решения прикладных задач, например, при создании бурятско-русского переводчика.

Следующий корпусный ресурс, называющийся «Диахронический корпус бурятского языка» (далее – ДКБЯ) [11], находится в стадии разработки. На данное время для ДКБЯ выполнена разметка письменно-монгольских словоформ объемом около 10 тыс. единиц, таблично сгруппированы аффиксы их словоизменения и составлен частотный словарь лексем на материале пяти бурятских летописей XIX в. (автор частотного словаря – О.С. Ринчинов), соответственно на вертикальной монгольской графике. Здесь следует отметить, что для ДКБЯ используется опубликованный письменно-монгольский текст в транслитерированном виде на латинице.

Описанные выше корпусы, как параллельный, диахронический можно считать по терминологии Е.В. Рахилиной [12, С. 14] специальными подкорпусами, в нашем случае, «Бурятского корпуса», хотя все данные ресурсы находятся на разных платформах, сайтах и серверах. Они напрямую между собой никак не связаны, между ними нет единой интеграции, как например, так называемые национальные корпуса, НКРЯ, British National Corpus, Национальный корпус калмыцкого языка и др. Такая ситуация сложилась ввиду того, что разработки по перечисленным корпусам бурятского корпуса выполнялись в разные годы отдельными независимыми друг от друга проектами с поддержкой также разных фондов. Характеризуя состояние бурятских корпусных ресурсов в целом, думается, что можно обозначить его как период становления. В силу эпизодичности финансовой поддержки в виде грантов / контрактов при их завершении, их заявленные цели и задачи соответственно выполняются, и разработка корпуса ставится на паузу. Такого рода повторяющиеся паузы приводят к отставанию / замедлению развития самого корпуса, в то время как технологии, инструментарии продолжают свое усовершенствование. Поэтому представляется крайне важной стабильная поддержка корпусного направления в бурятском языкознании в форме включения в проекты, реализующиеся в рамках государственных заданий. При стабильной финансовой поддержке будет соответственно расширяться и коллектив разработчиков, в особенности из среды молодых исследователей.

«Бурятский корпус» является письменным одноязычным и основным ресурсом из ряда вышеописанных. Наряду с его развитием, разрабатывается и корпус звукового формата, представляющий возможность получать информацию о звучании, просодии диалектов бурятского языка, отличающегося их разнообразием. Данный звуковой корпус будет являться специальным (см. выше). В отделе языкознания ИМБТ СО РАН плодотворно ведется работа над названным видом корпуса бурятского языка [13], [14]. Разработка звукового корпуса бурятских диалектов значительно обогатит представленность в киберпространстве языка бурят вместе с их речью, расширив технологические возможности для пользователей, как для её углубленного изучения или обучения, так и для ознакомления интересующимися, культурой одного из монголоязычных народов.

Методы и приемы лингвистического анализа корпусной и традиционной лингвистики, дополняя друг друга, определенным образом также совпадая, позволят получать совершенно новые результаты, выявить в бурятском языке вербальные явления, которые при традиционном анализе получить было невозможно. Методология корпусной лингвистики включает в себя автоматическую обработку текста, иначе - АОТ, как комплекс взаимосвязанных методов, приемов, процедур, начиная с предварительных автоматизированных поисков и извлечений искомых языковых данных из самих корпусов, как правило, с большими массивами текстовых, эмпирических данных, для их дальнейшего анализа. Использованием корпусных данных активизируются по-новому количественный / статистический, контекстного анализа, с применением, например, конкордансов, методы, предоставляя возможность для более глубокого описания, исследуемого языкового аспекта. В «Бурятском корпусе» имеется множество приемов отбора языковых данных, описание которых доступно в его настройках. Например, поиск лексических единиц можно осуществлять по словоформе, по лемме, а также – либо по всему ресурсу, либо - по текстовым источникам конкретного автора.

Каждый корпус имеет большие перспективы для своего дальнейшего, поступательного развития. Параллельный корпус бурятского языка может дополнятся другими языковыми переводами, как с бурятского, так и, например, с монгольского, английского и других языков. Кроме этого, надо отметить, параллельный корпус должен дополняться не только художественными переводами, но и направлением филологических / подстрочных переводов. Их можно назвать симметричными переводами. Все корпусные ресурсы востребованы не только в исследовательской деятельности, равным образом они могут использоваться в системе образования, в преподавании, обучении и изучении соответствующего языка. Не является исключением в связи с вышесказанным и параллельный корпус

бурятского языка. Помимо сказанного, текстовые базы данных в настоящее время крайне востребованы для проекта разработки автоматического бурятско-русского переводчика, о первичной онлайн версии которого, сообщается 27.06.2024 [15]. Для указанного проекта требуется первоначально параллельный корпус объемом не менее 50-100 тысяч выровненных предложений (бурятско-русских, русско-бурятских). Относительно вышесказанного филологического / построчного переводов надо пояснить их разницу в сравнении с художественными переводами, в которых наблюдается асимметрия, то есть не соблюдаются границы оригинального предложения, как пропуск литературным переводчиком части или целого предложения, нескольких предложений, даже крупных отрывком, добавление от переводчика личного текста, перестановки последовательности авторского текста в переводе. Параллельные художественные тексты с подобными расхождениями представляют определенную сложность для их использования при разработках автоматического переводчика.

Обсуждение

Разработка «Бурятского корпуса» выполняется с 2006 г. в формате долгосрочного инициативного научного проекта при поддержке различных научных фондов и госструктуры. Финансовая поддержка была оказана Минобразования и науки Республики Бурятия, РГНФ, РФФИ, ФФЛИ, Программой фундаментальных исследований Президиума РАН «Корпусная лингвистика». Выполнение работ по дальнейшему развитию бурятского корпуса предполагается быть включенным в плановый проект отдела языкознания ИМБТ СО РАН соответственно с финансированием по Госзаданию с 2026 г.

Одной из главных составляющих любого языкового корпуса являются тексты, специальным образом подготовленные. Наряду с тем, что исполнителями проводились работы по сканированию, редактированию, заключались договоры с официальными бурятскими издательствами по получению электронных версий текстов, также приобретались из открытого доступа, например, из СМИ и электронных библиотек. В конце XX в. – начале XXI в. книгопечатание и литературный процесс на бурятском языке переживали сложные времена, поэтому был сделан упор на художественную бурятскую литературу середины и II-й половины XX в. Тексты середины XX в. отражают соответственно в языковом аспекте свой вариант, в содержательном плане в основном, действительность того времени. Бурятский язык в произведениях, например, Х. Намсараева отличается от языка произведений современных авторов. Проблема репрезентативности текстовой составляющей бурятского корпуса стала решаться с постепенным налаживанием процесса книгопечатания на бурятском языке с 2010-х годов. Вместе с подготовкой текстовой составляющей параллельно решались вопросы программного обеспечения для ее интеграции в корпусный ресурс с соответствующими инструментариями, как интерфейс, осуществление настроек поиска, частотный и грамматический словари, последний с выполненной морфологической и мета- разметкой.

Заключение

Востребованность бурятских корпусных ресурсов неуклонно растет, что наблюдается по увеличению числа научных работ (статей, монографических исследований, диссертаций) по разным аспектам бурятского языка с использованием данных бурятского корпуса [16, С. 13-15]. Доступность онлайн языкового ресурса имеет большие плюсы, например, в условиях служебных поездок исследователей (командировки, экспедиции), не говоря уже об известных бывших ограничениях в периоды пандемии, поскольку языковой материал всегда доступен в режиме онлайн для продолжения сбора, дополнения или анализа и т.д.

В основных задачах развития бурятских корпусов всегда остаются планы углубления разработок, которые можно подразделить на два типа: 1. задачи общего характера, имеющие целью пополнение текстовых баз данных всех ресурсов в соответствии с принципами их репрезентативности и сбалансированности; 2. частные задачи по основному и специальным корпусам бурятского языка по отдельности. Частные задачи основного корпуса нацелены на разработку семантической разметки. По параллельному корпусу бурятского языка к задачам второго типа относятся продолжение полуавтоматического выравнивания бурятско-русских и обратных художественных текстов с последующим пополнением текстовой базы данных и морфологическая разметка текстов. Чем глубже и, чем больше разных видов разметок в корпусе, тем больше исследовательских, равно, как и образовательных, учебных, методических и тому подобных задач можно решать на его данных [17]. Лингвистическая разметка и/или аннотация, являясь и процессом, и его результатом, дает разнообразную информацию о текстовых материалах корпуса. Основными видами разметки «Бурятского корпуса» являются метаразметка и собственно лингвистическая разметка (в **Основных результатах** выше указано о разметке внешних данных текстов и морфологических признаках словоформ). Названые виды разметки, часто могут иметь в свою очередь свои подвиды (например, в «Бурятском корпусе» выдается частеречная принадлежность словоформ). Все виды работ по разрабатываемым корпусам, их программные составляющие выполняются в тесном сотрудничестве с соответствующими специалистами, как компьютерные, корпусные лингвисты и естественно, самими языковедами, специализирующимися в области бурятского, в целом, монгольского языкознания.

Финансирование

Работа выполнена в рамках государственного задания (проект «Мир человека в монгольских языках: анализ средств выражения эмотивности»).

Благодарности

Институт монголоведения, буддологии и тибетологии СО РАН, Улан-Удэ, Российская Федерация

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Funding

The work was carried out within the framework of a state assignment (the project “The Human World in the Mongolian Languages: Analysis of Means of Expressing Emotivity”).

Acknowledgement

Institute for Mongolian, Buddhist and Tibetan Sciences, Siberian Branch, Russian Academy of Sciences, Ulan-Ude, Russian Federation

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Национальный корпус русского языка стал одним из самых востребованных инструментов русистов во всем мире.— URL: https://www.ras.ru/news/shownews.aspx?id=15adbfc1-d29c-483c-b983-fbe63ca3e110 (дата обращения: 15.04.2024).

2. I-я стратегическая сессия «Информационные технологии и языки народов России» // I-я стратегическая сессия «Информационные технологии и языки народов России». — 2024 — URL: https://fadn.gov.ru/press-centr/news/i-strategicheskaya-sessiya-%C2%ABinformaczionnyie-texnologii-i-yazyiki-narodov-rossii%C2%BB (дата обращения: 15.05.2024)

3. Дырхеева Г.А. Конкордансы, словоуказатели (индексы) и частотные словари к литературным произведениям / Г.А. Дырхеева. — Улан-Удэ, 1981. — 27 с. — Деп. в ИНИОН АН СССР. — 28.05.82, № 10294.

4. Дырхеева Г.А. Использование частотного словаря для оптимизации преподавания бурятского языка. / Г.А. Дырхеева — Улан-Удэ: Изд-во БНЦ СО РАН, 1992. — 238 с.

5. Корпус бурятского языка. — 2011 — URL: http://corpora.imbtarchive.ru/index.php (дата обращения: 25.04.2024)

6. Бурятский корпус. — 2012 — URL: http://web-corpora.net/BuryatCorpus/search/?interface_language=ru (дата обращения: 25.04.2024)

7. Шагдаров Л.Д. Бурятско-русский словарь / Л.Д. Шагдаров, К.М. Черемисов — Улан-Удэ: ОАО «Республиканская типография», 2006. — 636 с.

8. Шагдаров Л.Д. Бурятско-русский словарь / Л.Д. Шагдаров, К.М. Черемисов — Улан-Удэ: ОАО «Республиканская типография», 2008. — 708 с.

9. Архангельский Т.А.. Корпусная платформа Tsakorpus и языки России / Т.А. Архангельский // Электронная письменность народов Российской федерации – 2021 & IWCLUL 2021. Материалы Междунар. научно-практич. конф.; — Сыктывкар: Изд-во «Коми республиканская академия государственной службы и управления», 2021. — с. 23-24.

10. Параллельный бурятско-русский корпус. — 2016 — URL: https://ruscorpora.ru/new/search-para.html?lang=bua (дата обращения: 25.04.2024)

11. Диахронический корпус бурятского языка // Диахронический корпус бурятского языка. — 2020 — URL: http://annals.imbtarchive.ru/ (дата обращения: 25.04.2024)

12. Рахилина Е.В.. Корпус как творческий процесс / Е.В. Рахилина // Национальный корпус русского языка: 2006-2008. — Новые результаты и перспективы.; — Санкт-Петербург: СПб : Несто-История, 2009. — с. 7-25.

13. Абаева Л.Д.. Вопросы разработки звукового корпуса бурятских диалектов / Л.Д. Абаева // Предложение как единица речи. Материалы Всероссийского научного симпозиума с международным участием, посвященного 95-летию со дня рождения М.И. Черемисиной. — Новосибирск: Изд-во ФГБУН Институт филологии, 2019. — с. 165-167.

14. Абаева Ю.Д. Геоинформационный веб-ресурс «Диалектный корпус бурятского языка / Ю.Д. Абаева, О.С. Ринчинов // Филологические науки. Вопросы теории и практики. — 2023. — 1. — с. 328-334.

15. Google Translate (Google Переводчик) получил самое большое обновление – добавилось 110 языков, включая языки регионов России. — 2024. — URL: https://www.ixbt.com/news/2024/06/27/google-translate-google-110.html (дата обращения: 01.07.2024)

16. Скрибник Е.К. Подлежащее в бурятских конструкциях каузации эмоций / Е.К. Скрибник, Н.Д. Даржаева // Языки и фольклор коренных народов Сибири. — 2024. — 1(49). — с. 9-23.

17. Толдова С.Ю. Разметка лингвистическая / С.Ю. Толдова, Е.А. Логинова, Д.П. Попова // Разметка лингвистическая. — 2011 — URL: https://clck.ru/3Bm726 (дата обращения: 01.07.2024)

Список литературы на английском языке / References in English

1. Natsional'nyj korpus russkogo jazyka stal odnim iz samyh vostrebovannyh instrumentov rusistov vo vsem mire [The National Corpus of the Russian Language has become one of the most sought-after tools for Russianists around the world]. — URL: https://www.ras.ru/news/shownews.aspx?id=15adbfc1-d29c-483c-b983-fbe63ca3e110 (accessed: 15.04.2024) [in Russian]

2. I-ja strategicheskaja sessija «Informatsionnye tehnologii i jazyki narodov Rossii» [1st strategic session "Information technologies and languages of the peoples of Russia"] // 1st strategic session "Information technologies and languages of the

peoples of Russia". — 2024 — URL: https://fadn.gov.ru/press-centr/news/i-strategicheskaya-sessiya-%C2%ABinformaczionnyie-texnologii-i-yazyiki-narodov-rossii%C2%BB (accessed: 15.05.2024) [in Russian]

3. Dyrkheeva G.A. Konkordansy, slovoukazateli (indeksy) i chastotnye slovari k literaturnym proizvedeniyam [Concordances, word indexes and frequency dictionaries for literary works] / G.A. Dyrkheeva. — Ulan-Ude, 1981. — 27 p. — Dep. in INION AN SSSR. — 28.05.82, № 10294 [in Russian].

4. Dyrheeva G.A. Ispol'zovanie chastotnogo slovarja dlja optimizatsii prepodavanija burjatskogo jazyka. [Using a frequency dictionary to optimize teaching of the Buryat language] / G.A. Dyrheeva — Ulan-Ude: BNTs SO RAN Publishing, 1992. — 238 p. [in Russian]

5. Korpus burjatskogo jazyka [Corpus of the Buryat language]. — 2011 — URL: http://corpora.imbtarchive.ru/index.php (accessed: 25.04.2024) [in Russian]

6. Burjatskij korpus [Buryat Corpus]. — 2012 — URL: http://web-corpora.net/BuryatCorpus/search/?interface_language=ru (accessed: 25.04.2024) [in Russian]

7. Shagdarov L.D. Burjatsko-russkij slovar' [Buryat-Russian dictionary] / L.D. Shagdarov, K.M. Cheremisov — Ulan-Ude: OAO «Respublikanskaja tipografija», 2006. — 636 p. [in Russian]

8. Shagdarov L.D. Burjatsko-russkij slovar' [Buryat-Russian dictionary] / L.D. Shagdarov, K.M. Cheremisov — Ulan-Ude: OAO «Respublikanskaya tipografiya», 2008. — 708 p. [in Russian]

9. Arhangel'skij T.A.. Korpusnaja platforma Tsakorpus i jazyki Rossii [Tsakorpus Corpus Platform and Russian Languages] / T.A. Arhangel'skij // Electronic Writing of the Peoples of the Russian Federation – 2021 & IWCLUL 2021. Proc. of the Int. scientific-practical. conf.; — Syktyvkar: «Komi respublikanskaja akademija gosudarstvennoj sluzhby i upravlenija» Publishing, 2021. — p. 23-24. [in Russian]

10. Parallel'nyj burjatsko-russkij korpus [Parallel Buryat-Russian Corpus]. — 2016 — URL: https://ruscorpora.ru/new/search-para.html?lang=bua (accessed: 25.04.2024) [in Russian]

11. Diahronicheskij korpus burjatskogo jazyka [Diachronic corpus of the Buryat language] // Diachronic corpus of the Buryat language. — 2020 — URL: http://annals.imbtarchive.ru/ (accessed: 25.04.2024) [in Russian]

12. Rahilina E.V.. Korpus kak tvorcheskij protsess [The Corpus as a Creative Process] / E.V. Rahilina // National Corpus of the Russian Language: 2006-2008. — New Results and Prospects.; — Sankt-Peterburg: SPb : Nesto-Istorija, 2009. — p. 7-25. [in Russian]

13. Abaeva L.D.. Voprosy razrabotki zvukovogo korpusa burjatskih dialektov [Issues of development of the sound corpus of Buryat dialects] / L.D. Abaeva // Sentence as a unit of speech. Proceedings of the All-Russian scientific symposium with international participation dedicated to the 95th anniversary of M.I. Cheremisina's birthday. — Novosibirsk: FGBUN Institut filologii Publishing, 2019. — p. 165-167. [in Russian]

14. Abaeva Ju.D. Geoinformatsionnyj veb-resurs «Dialektnyj korpus burjatskogo jazyka [Geoinformation web resource "Dialect corpus of the Buryat language] / Ju.D. Abaeva, O.S. Rinchinov // Philological sciences. Theoretical and practical issues. — 2023. — 1. — p. 328-334. [in Russian]

15. Google Translate (Google Perevodchik) poluchil samoe bol'shoe obnovlenie – dobavilos' 110 jazykov, vkljuchaja jazyki regionov Rossii [Google Translate (Google Translater) received the biggest update — 110 languages were added, including languages of Russian regions]. — 2024 — URL: https://www.ixbt.com/news/2024/06/27/google-translate-google-110.html (accessed: 01.07.2024) [in Russian]

16. Skribnik E.K. Podlezhaschee v burjatskih konstruktsijah kauzatsii emotsij [Subject in Buryat constructions of causation of emotions] / E.K. Skribnik, N.D. Darzhaeva // Languages and folklore of the indigenous peoples of Siberia.. — 2024. — 1(49). — p. 9-23. [in Russian]

17. Toldova S.Ju. Razmetka lingvisticheskaja [Linguistic markup] / S.Ju. Toldova, E.A. Loginova, D.P. Popova // Linguistic markup. — 2011 — URL: https://clck.ru/3Bm726 (accessed: 01.07.2024) [in Russian]