

RU

Геоинформационный веб-ресурс «Диалектный корпус бурятского языка»

Ринчинов О. С., Абаева Ю. Д.

Аннотация. Цель исследования - обеспечить открытый доступ к структурированным и аннотированным звуковым данным диалектного корпуса бурятского языка. Представление корпуса в Сети было решено выполнить в виде геоинформационной системы с привязкой данных к цифровой карте, так как при классификации бурятских диалектов территориальный принцип играет одну из ведущих ролей. Первоначально была составлена программа речевого корпуса и получены звукозаписи в исполнении дикторов - носителей диалектов. Записанный материал был подвергнут сегментации и аннотированию в программе ELAN. Следующим шагом была разработка программы, позволяющей перенести данные из файлов форматов ELAN в реляционную базу данных. Для представления данных в Интернете было разработано веб-приложение в виде интерактивной цифровой карты на основе Google Maps Platform. В результате был создан веб-ресурс, который предоставляет пользователям доступ к звуковым диалектным данным, представленным в аннотированном и структурированном виде и отображаемым по географическому принципу. Научная новизна заключается во вводе в научный и публичный оборот материалов принципиально нового вида, позволяющих получить информацию о современном звучании бурятских диалектов, а также проводить исследования современной бурятской диалектной речи.

EN

Geoinformation Web Resource “The Dialect Corpus of the Buryat Language”

Rinchinov O. S., Abaeva I. D.

Abstract. The study aims to provide open access to structured and annotated sound data of the dialect corpus of the Buryat language. It was decided to present the corpus on the Web in the form of a geoinformation system with data binding to a digital map, since the territorial principle plays one of the leading roles in the classification of Buryat dialects. First, the programme of the speech corpus was compiled and sound recordings performed by informants - speakers of the dialects were obtained. The recorded material was segmented and annotated in the ELAN programme. The next step was to develop a programme that allows transferring data from ELAN format files to a relational database. To present data on the Internet, a web application was developed in the form of an interactive digital map based on Google Maps Platform. As a result, a web resource was created that provides users with access to audio dialect data presented in an annotated and structured form and displayed according to the geographic principle. Scientific novelty lies in introducing into scientific and public use materials of a fundamentally new type that make it possible to obtain information about the modern sound of Buryat dialects, as well as to conduct research on modern Buryat dialect speech.

Введение

Бурятский язык распространен на огромной территории в Восточной Сибири, протянувшейся почти на полторы тысячи километров с запада на восток. Ввиду исторической дисперсности и изолированности расселения бурятских субэтнических групп для этого языка характерна многодиалектность, обусловленная рядом факторов, сохраняющаяся и поныне (Дырхеева, 2014, с. 199). Поскольку диалекты в настоящее время являются основной формой живого функционирования бурятского языка, острой является проблема их изучения.

Устный диалектный корпус бурятского языка создавался в рамках проекта «Диалектный корпус как средство сохранения языкового разнообразия», поддержанного грантом РФФИ. Целью проекта является создание звукового корпуса, отражающего синхронное состояние бурятской диалектной речи. Материал корпуса содержит

как специально составленные высказывания различной сложности и коммуникативной направленности, так и образцы спонтанной речи (монологи на повседневные темы: автобиография, праздники, исторические предания, интересные события и т.д.). Корпус в первую очередь нацелен на исследование просодических особенностей бурятских диалектов, поэтому записанный материал был подвергнут сегментированию и просодической разметке.

Одной из основных целей создания такого ресурса было предоставление доступа к диалектной звуковой материи языка для пользователей Интернета. Актуальность такой разработки заключается в необходимости сохранения современного состояния бурятского языка, его популяризации, предоставления материала для диалектных лингвистических исследований. Звучащий текст, подвергнутый просодической разметке, станет основой для изучения бурятской диалектной речи не только на суперсегментном уровне, но и на лексическом, синтаксическом, морфологическом. В связи с этим перед исследователями ставились следующие задачи:

- разработать программный инструмент для автоматизированного преобразования файлов, созданных программой ELAN в результате обработки записанного речевого материала, в таблицу с последующим импортом в реляционную базу данных;
- разработать веб-приложение для представления Диалектного корпуса бурятского языка в Интернете в виде геоинформационной системы.

Методы исследования включают декомпозицию основных задач, классификацию и моделирование предметной области, алгоритмизацию найденных решений и другие подходы системного анализа.

Теоретической базой работы послужили принципы и подходы, сложившиеся в таких областях, как корпусная лингвистика, обработка мультимедийных данных, геоинформационные системы.

В работе над проектом мы опирались на имеющийся в российской лингвистике опыт создания устных корпусов, представляющих региональные варианты языков и диалектов, на труды, посвященные описанию принципов составления корпусов и представления их в Сети (Бондарко, Скредин, Вольская и др., 2000; Рассказы о сновидениях..., 2009; Andriyanets, Daniel, Pakendorf, 2018; Kachkovskaya, Kocharov, Skrelin et al., 2016).

Быстро развивающейся областью применения информатики в гуманитарных науках является цифровая картография (Владимиров, 2005). Так, в нашем случае она позволяет исследовать пространственное измерение диалектного многообразия, а в качестве инструмента выступает особый класс программного обеспечения, называемый географическими информационными системами (ГИС) (Жданова, Бельх, 2014). Применение цифровых карт существенно расширяет методическую базу исследований, а богатые возможности агрегации и визуализации данных, которыми обладают современные картографические сервисы, существенно расширяют область применения этого инструмента и позволяют рассматривать ГИС как удобное и продуктивное аналитическое средство (Ватлина, Лунькова, 2015).

Практическая значимость работы заключается в том, что открытый доступ к специальным образом структурированным и аннотированным звуковым данным корпуса обеспечивает ценную, ранее не доступную источникную базу для ученых, преподавателей, студентов, занимающихся изучением бурятских диалектов, а также для широкого круга людей, интересующихся бурятским языком.

Основная часть

Материал Диалектного корпуса бурятского языка

Работа над корпусом состояла из следующих этапов: Разработка программы корпуса → Запись речевого материала → Сегментирование и аннотирование звуковых файлов в программе ELAN → Преобразование файлов ELAN в базу данных → Представление корпуса в виде веб-ресурса.

Программа сбора материала была составлена таким образом, чтобы наиболее ярко выделить характерные интонационные особенности диалектов. Первая часть – чтение заранее подготовленных фраз, отражающих различные коммуникативные типы высказываний, различной длительности и синтаксической структуры. Так, однословное повествовательное высказывание *Үбэл* 'Зима', постепенно расширяясь: *Хүйтэн үбэл* 'Холодная зима'; *Хүйтэн үбэл байгаа* 'Была холодная зима'; *Нёдондо хүйтэн үбэл байгаа* 'В прошлом году была холодная зима' – позволяет проследить изменения интонации завершенности в связи с увеличением длительности высказывания, особенностей его коммуникативной структуры внутри диалекта, а также позволяет провести сравнительный анализ интонации различных диалектов на однотипном материале.

Вторая часть программы – продуцирование дикторами спонтанного монологического текста на заданную тему. Были предложены следующие темы повествований: автобиография; рассказы о семье, профессии, праздниках, интересных событиях; фольклорные тексты (сказки, легенды, предания). Для сравнения диалектных интонационных особенностей в первую очередь внимание уделялось автобиографическим текстам, так как они имеют сходную структуру и однообразное интонационное оформление.

Запись материала проводилась с привлечением дикторов, с детства владеющих родным диалектом и использующих его в качестве средства повседневного общения. При выборе диктора, помимо его половозрастных характеристик, обязательным параметром был язык его повседневного общения и проживание в сельской местности, на так называемой «малой Родине». Проживание в однородной языковой среде, общение преимущественно с носителями своего диалекта позволяют сохранить произносительные диалектные особенности. Указание на место жительства информанта явилось ведущим фактором при представлении материала корпуса в Сети.

Выбор представления корпуса в виде геоинформационной системы с привязкой к географической карте был обусловлен тем, что при классификации бурятских диалектов территориальный принцип играет весомую роль. Помимо структурных особенностей диалектов: фонетико-грамматических или лексико-синтаксических – при их классификации большое значение имеют такие факторы, как принадлежность его носителей к тому или иному родоплеменному объединению (эхиритаге, булагатам, хоринцам, сартулам, сонглолам, аларцам и др.), которая имеет географическую привязку. Родоплеменной принцип классификации во многом совпадает с территориальным делением бурятских диалектов на восточные (восточное побережье оз. Байкал), западные (западное побережье оз. Байкал) и южные с некоторыми исключениями. Помимо этих групп имеется группа говоров, которая по территориальному признаку обозначена как присаянская (Рассадин, 1996) или по родоплеменному признаку – как аларо-тункинская (Бураев, 1996, с. 14). Также имеются два говора, не вошедшие ни в одну из диалектных групп, они рассматриваются в качестве изолированных самостоятельных говоров и по территориальному признаку их можно назвать крайним западным – нижнеудинский и крайним восточным – хамниганский говор (Бураев, 1996, с. 14).

На сегодняшний день материал корпуса включает записи большинства диалектов бурятского языка. Так, восточная группа говоров в корпусе представлена литературным бурятским языком, основанным на хоринском диалекте (дикторами выступили представители Хоринского и Кижингинского районов Республики Бурятия), а также агинским подговором хоринского диалекта (пос. Агинское, Забайкальский край). В Таблице 1 приводятся территориальные группы говоров, представленные в корпусе, с информацией о месте жительства дикторов – носителей диалектов.

Таблица 1. Территориальное распределение дикторов – носителей диалектов

Территориальная группа говоров	Говор	Место жительства диктора
Восточные	хоринский	с. Кижинга, Кижингинский р-н Республики Бурятия (РБ), с. Хоринск, Хоринский р-н РБ
	агинский подговор	пос. Агинское, Забайкальский край
Западные	эхирит-булагатский	с. Баяндай, Баяндаевский р-н, Иркутская область
	боханский	с. Хохорск, Боханский р-н, Иркутская обл.
	баргузинский	с. Баянгол, Баргузинский р-н РБ
Южные	цонгольский	г. Кяхта, Кяхтинский р-н РБ
	сартульский	с. Петропавловка, Джидинский р-н РБ
Присаянские	закаменский	с. Санага, Закаменский р-н РБ
	тункинский	пос. Аршан, с. Галбай, Тункинский р-н РБ
Изолированный диалект (крайний западный)	нижнеудинский	с. Кушун, Нижнеудинский р-н, Иркутская область

Формирование базы данных Диалектного корпуса бурятского языка на основе аннотаций, выполненных в программе ELAN

Записанный материал Диалектного корпуса бурятского языка был подвергнут сегментации и аннотированию в специализированной программе разметки мультимедийных данных ELAN, созданной в Институте психолингвистики Макса Планка (Wittenburg, Brugman, Russel et al., 2006; ELAN (Version 6.4). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive, 2022. URL: <https://archive.mpi.nl/tla/elan>).

Она предлагает следующие основные функции:

- создание множества иерархически организованных слоев аннотирования (*tiers*, согласно терминологии программы ELAN);
- синхронизация аннотаций с временной шкалой источника;
- поддержка контролируемых словарей (авторитетного контроля) (Pennington, 2014).

Примененная к звукозаписям бурятского диалектного корпуса, схема аннотирования выглядит следующим образом, см. Таблицу 2.

Таблица 2. Схема уровней аннотирования звуковых данных диалектного корпуса

Название	Содержание
Text	Орфографическая расшифровка текста.
Phrase	Орфографическая расшифровка текста с делением на синтагмы.
Punct	Обозначение иллокутивной функции синтагмы: завершенность (.), незавершенность (,), вопросительность (?), восклицательность (!) и т.п.
Pause	Паузы абсолютные (А) и заполненные (э-э, м-м).
Word	Орфографическая запись текста с сегментацией на слова.
Accent	Выраженное изменение тона: / – восходящее, \ – нисходящее, ^ – восходяще-нисходящее, v – нисходяще-восходящее, – – ровное.
Level	Высота тона в терминах автосегментной школы (Н, М, L). Указывается на начале, конце и интонационном центре синтагмы, а также на участках выраженного изменения тона.
Неречевые явления	Например, смех {laugh}.
Tempo	Замедление (len) и ускорение (fast) темпа.
Loudness	Повышение (loud) и понижение (silent) громкости.

Для каждого обработанного в программе мультимедийного источника ELAN генерирует файлы двух типов, соответствующие спецификациям XML: файл служебной информации PFSX, в котором описываются слои аннотирования, и файл данных EAF, в котором сохраняются результаты аннотирования. Преобразование аннотаций речевых записей, представленных в диалектном корпусе, из формата XML в табличный вид и последующее размещение в реляционной базе данных позволяют применять разнообразные способы анализа размеченных звуковых данных, например статистическую обработку ритмико-интонационных (просодических) характеристик речи и пр. (Кривнова, Смирнова, 2019, с. 324). Кроме того, перенос данных в системе управления базами данных (СУБД) Oracle MySQL создает удобную основу для разработки веб-приложения для Диалектного корпуса бурятского языка.

Для этих целей на языке программирования Visual Basic for Applications был реализован программный инструмент для автоматизированного преобразования файлов разметки диалектных аудиозаписей, создаваемых с помощью программы ELAN и имеющих расширение EAF, из формата XML в табличное представление и последующий импорт табличных данных в СУБД Oracle MySQL. При этом полностью сохраняются аннотации звуковых данных (содержание уровней аннотирования).

VBA-скрипт сначала анализирует и индексирует временные интервалы, сформированные при аннотировании звуковых данных, представленных в файле аудиозаписи. Затем он извлекает размеченные данные из соответствующих уровней аннотирования («слоев») и размещает их в строке по индексу соответствующего временного интервала. Поскольку XML-структура каждого слоя в примененной схеме разметки однородна, скрипт позволяет обрабатывать произвольные уровни.

Извлеченные из файла разметки EAF табличные данные построчно атрибутируются идентификатором соответствующей аудиозаписи и размещаются в реляционной базе данных, что обеспечивает удобство их последующих обработки и анализа.

Геоинформационная система «Диалектный корпус бурятского языка»

При обработке и представлении диалектных материалов важно сохранить информацию об обстоятельствах их фиксации – времени, месте, информанте, собирателе и т.п. Необходимо также представить ряд физических параметров аудиозаписей, данные об их инвентаризации и хранении. Информация для атрибутирования первичных материалов по этим параметрам обычно содержится в звуковых аннотациях, экспедиционных дневниках и собирается при первичной обработке материалов (см. Таблицу 3), чтобы предоставить возможность однозначно идентифицировать диалектную аудиозапись и определить ее положение как в пространственно-временном аспекте, так и по отношению к идиомам бурятского языка.

Таблица 3. Схема учета диалектных аудиозаписей

Параметр описания произведения	Содержание
ОСНОВНЫЕ СВЕДЕНИЯ	
Заголовок	Заголовок или контекст записи
Диалект (идиом)	Указывается диалект, к которому относится запись
ПАСПОРТ ЗАПИСИ	
Собиратель(и)	ФИО собирателей
Место записи	Указывается географическая локализация, в которой была выполнена запись
Время записи	Указывается полная или неполная дата, когда была осуществлена запись
ПАСПОРТ ИНФОРМАНТА	
Информант	ФИО информанта
Год рождения	Год рождения информанта
Место рождения	Место рождения информанта
Место проживания	Место постоянного проживания информанта
Родовая принадлежность	Родоплеменная принадлежность информанта
Образование	Образование информанта (среднее, неоконченное высшее, высшее, ученая степень)
Язык	Язык и диалект постоянного общения
АРХИВНЫЕ ССЫЛКИ	
Код	Учетный номер
Файл аудиозаписи	Ссылка на исходный файл, полученный с цифрового записывающего устройства
Файл обработанной аудиозаписи	Ссылка на файл, созданный после обработки исходной аудиозаписи

Для создания онлайн-ресурса, нацеленного на представление аудиообразцов бурятских диалектов, в качестве основного был выбран картографический способ агрегации и отображения информации. Это объясняется тем, что идиомы бурятского языка, как правило, географически локализованы, а место записи является обязательным атрибутом, который при обработке материала определяется в первую очередь. Современные онлайн-геоинформационные сервисы предоставляют обширные возможности для агрегирования и вывода различной информации, связанной с отображаемыми на цифровой карте объектами. Таким образом,

картографический подход позволяет визуализировать географический размах социолингвистического явления, явным образом указать локализацию тех или иных идиомов в пространстве этнической Бурятии.

Информационная модель веб-ресурса была реорганизована с учетом требований авторитетного контроля. В качестве объектов авторитетного контроля выступили следующие основные типы объектов (классы): (1) лицо (person); (2) место (place), представленные в соответствующих таблицах реляционной базы данных.

В основу описания указанных классов были положены функциональные требования к авторитетным записям (FRAD), описанные в международном стандарте (Функциональные требования..., 2011). В реализации базы данных диалектного звукового корпуса атрибуты, установленные FRAD для объектов авторитетного контроля, были использованы частично. Полное заполнение атрибутивной части авторитетного файла составляет самостоятельную исследовательскую задачу, находящуюся вне целей настоящего исследования.

Основным объектом реляционной базы данных является «звукозапись» (record), объединяющая в себе разделы «Основные сведения», «Паспорт записи» и «Архивные ссылки» из первичного описания диалектной аудиозаписи (см. Таблицу 3). Она содержит также ссылки на файлы, в которых содержатся цифровая запись (или фрагмент записи) и результаты разметки звукового источника в программе ELAN. Для отношения «record – place» была реализована связь «один – к одному». «Звукозапись» связана с «лицами» (person) отношением «действующее лицо» (actor), в каждом элементе которого указывается код записи, код лица и его роль по отношению к записи – «информант» или «собиратель». При этом для отношения «record – actor» реализована связь «один – ко многим». Для определения важных в контексте представления диалектологических материалов параметров была выделена категория «идиом» (idiom), объекты которой представлены в соответствующей таблице. Определено прямое одноместное отношение «record – idiom».

Выполненная на основе указанных принципов схема базы данных дает возможность реализовать следующие способы агрегирования информации о диалектных аудиозаписях: относительно места записи, относительно информанта, относительно собирателя, относительно идиома. Первый из перечисленных способов организации информации позволяет реализовать картографическое отображение данных.

Для картографического представления информации был выбран геоинформационный сервис Google Maps, предлагающий интерфейс прикладного программирования на языке JavaScript, с помощью которого цифровая карта легко интегрируется в функционал веб-сайта (Google Maps Tutorial. W3Schools. URL: https://www.w3schools.com/graphics/google_maps_intro.asp).

Backend веб-ресурса реализован на основе СУБД Oracle MySQL, запросы к базе данных и их обработка выполнялись с помощью AJAX-запросов, оформленных на языке программирования PHP (Soni S. How to Use AJAX in PHP and jQuery. Envato Tuts+. 2020. URL: <https://code.tutsplus.com/tutorials/how-to-use-ajax-in-php-and-jquery--cms-32494>). Таким образом, веб-приложение полностью реализовано на свободно распространяемом программном обеспечении. Оно размещено на портале Центра восточных рукописей и ксилографов Института монголоведения, буддологии и тибетологии Сибирского отделения Российской академии наук (URL: <http://dialectology.imbtarchive.ru>).

Рабочая зона веб-приложения состоит из двух интерактивных панелей, на одной из которых размещается электронная карта Google Maps с выведенными на ней объектами типа «record», вторая служит для вывода различной рабочей информации. На карте отображаются места сбора аудиозаписей. При активизации пользователем маркера объекта Google Maps во всплывающем инфоокне выводятся следующие сведения о соответствующей аудиозаписи, оформленные как ссылки на отображение связанных с ней данных из базы данных веб-приложения:

- 1) название (контекст) диалектологической аудиозаписи;
- 2) идиом;
- 3) место фиксации аудиозаписи;
- 4) информант;
- 5) собиратель.

Когда пользователь активирует ссылку по названию аудиозаписи, на информационной панели выводится полная информация о выбранном произведении и отображается панель программного аудиопроигрывателя, с помощью которого можно прослушать аудиозапись, представленную в формате MP3. Если запись обработана в программе ELAN, выводится ссылка на табличное отображение выполненных для нее аннотаций (см. предыдущий раздел), которое открывается в отдельной вкладке веб-браузера. Активация других ссылок запускает скрипты веб-приложения, которые выдают списки аудиозаписей, агрегированные по указанному значению соответствующего поискового запроса. Таким образом можно получить полный перечень записей, относящихся к выбранному идиому или записанных в определенном месте. Применяя разные способы отображения данных, пользователь может составить полное представление о диалектах бурятского языка, изученных исследователями в ходе выполнения проекта.

Заключение

В результате реализации текущего этапа проекта «Диалектный корпус бурятского языка как средство сохранения языкового разнообразия» было осуществлено преобразование звукового корпуса в удобный для выполнения

аналитических исследований формат. Созданная на основе просодического аннотирования диалектных аудиозаписей реляционная база данных позволяет применять разнообразные методы изучения корпусных данных, включая лингвостатистические. Разработанная геоинформационная система «Диалектный корпус бурятского языка» открывает доступ к значительному объему обработанных материалов звукового корпуса, которые организованы по географическому принципу, ключевому в территориальном распределении диалектов бурятского языка. Таким образом, в распоряжение пользователей ГИС предоставляется ценная информация, позволяющая составить полное представление о диалектах бурятского языка, исследованных в ходе реализации проекта.

Перспективы дальнейшего исследования проблемы мы видим в развитии инструментов анализа корпусных данных, в первую очередь лингвостатистических, расширении речевого материала, в более детальном аннотировании, включающем указания на морфологические, лексические и синтаксические характеристики. Разработанные в ходе выполнения проекта программные средства и инструменты формируют открытую платформу для создания звуковых корпусов и их эффективного исследования.

Финансирование | Funding

RU Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-012-00491.

EN The reported study was funded by the RFBR as a part of research project No. 20-012-00491.

Источники | References

1. Бондарко Л. В., Скредин П. А., Вольская Н. Б., Шерстинова Т. Ю. Электронные звуковые коллекции в Интернет // Электронные библиотеки. 2000. Т. 3. № 1.
2. Бураев И. Д. Проблемы классификации бурятских диалектов // Проблемы бурятской диалектологии: сб. ст. Улан-Удэ: Изд-во Бурятского научного центра Сибирского отделения Российской академии наук, 1996.
3. Ватлина Т. В., Лунькова Е. С. Пространственный анализ диалектного словообразовательного типа, функционирующего на территории Смоленской области // ИнтерКарто. ИнтерГИС. 2015. Т. 21.
4. Владимиров В. Н. История и география: пути взаимодействия // Вестник Новосибирского государственного университета. Серия «История, филология». 2005. Т. 4. Вып. 2. История.
5. Дырхеева Г. А. Литературный бурятский язык: история и проблематика // Вестник Бурятского научного центра Сибирского отделения Российской академии наук. 2014. № 1 (13).
6. Жданова Е. А., Белых А. А. Географические информационные системы в лингвистических исследованиях // Интеллектуальные системы в производстве. 2014. № 2 (24).
7. Кривнова О. Ф., Смирнова О. С. Интроспективная просодическая разметка письменного текста и его реальное озвучивание (сравнительный анализ на материале коллекции текстов Р. И. Аванесова) // Компьютерная лингвистика и интеллектуальные технологии: по мат. ежегод. междунар. конф. «Диалог» (г. Москва, 29 мая - 1 июня 2019 г.). М., 2019. Вып. 18 (25).
8. Рассадин В. И. Присяянская группа бурятских говоров. Улан-Удэ: Изд-во Бурятского научного центра Сибирского отделения Российской академии наук, 1996.
9. Рассказы о сновидениях: корпусное исследование устного русского дискурса / под ред. А. А. Кибрика и В. И. Подлесской. М.: Языки славянских культур, 2009.
10. Функциональные требования к авторитетным данным: концептуальная модель: заключительный отчет, декабрь 2008 / под ред. Г. Е. Патона; пер. с англ. О. А. Лаврёновой. СПб.: Российская национальная библиотека, 2011.
11. Andriyanets V., Daniel M., Pakendorf B. Discovering Dialectal Differences Based on Oral Corpora // Компьютерная лингвистика и интеллектуальные технологии: по мат. ежегод. междунар. конф. «Диалог» (г. Москва, 30 мая - 2 июня 2018 г.) / под общ. ред.: В. Селегей, И. М. Кобозева, Т. Е. Янко, И. Богуславский, Л. Л. Иомдин, М. А. Кронгауз, А. Ч. Пиперски. М.: Российский государственный гуманитарный университет, 2018. Вып. 17 (24).
12. Kachkovskaya T. V., Kocharov D. A., Skrelin P. A., Volskaya N. B. CoRuSS - a New Prosodically Annotated Corpus of Russian Spontaneous Speech // Proceedings of the Tenth Conference on International Language Resources and Evaluation (LREC'16) (Portorož, May 23-28, 2016). Portorož: European Language Resources Association, 2016.
13. Pennington R. Producing Time-Aligned Interlinear Texts: Towards a SayMore-FLEEx-ELAN Workflow. 2014. URL: https://www.academia.edu/6474779/Producing_time_aligned_interlinear_texts_Towards_a_SayMore_FLEEx_ELAN_workflow
14. Wittenburg P., Brugman H., Russel A., Klassmann A., Sloetjes H. ELAN: A Professional Framework for Multimodality Research // Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation. 2006. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf

Информация об авторах | Author information**RU****Ринчинов Олег Сергеевич¹**, д. ист. н.**Абаева Юлия Догоржаповна²**, к. филол. н.^{1,2} Федеральное государственное бюджетное учреждение науки

«Институт монголоведения, буддологии и тибетологии

Сибирского отделения Российской академии наук» (ИМБТ СО РАН), г. Улан-Удэ

EN**Rinchinov Oleg Sergeevich¹**, Dr**Abaeva Iuliia Dogorzhapovna²**, PhD^{1,2} Federal State Budgetary Institution of Science

"Institute for Mongolian, Buddhist and Tibetan Studies

of the Siberian Branch of the Russian Academy of Sciences", Ulan-Ude

^{1,2} julaba@yandex.ru**Информация о статье | About this article**

Дата поступления рукописи (received): 12.12.2022; опубликовано (published): 31.01.2023.

Ключевые слова (keywords): звуковой корпус; бурятский язык; диалект; аннотирование; геоинформационная система; sound corpus; Buryat language; dialect; annotation; geoinformation system.