

СОЗДАНИЕ ПАРАЛЛЕЛЬНОГО УСТНОГО МОНГОЛЬСКО-РУССКОГО КОРПУСА (НА МАТЕРИАЛЕ ХАЛХАССКОГО И ОЙРАТСКИХ ДИАЛЕКТОВ ЗАПАДНОЙ МОНГОЛИИ)
CREATION OF A PARALLEL MONGOLIAN-RUSSIAN SPEECH CORPUS BASED ON THE MATERIAL OF KHALHA-MONGOLIAN AND OIRAT DIALECTS OF WESTERN MONGOLIA

Власов М.С. /Ph.D/
АГППУ им В.М.Шукина, Россия,
Одончимэг Т. /Ph.D/
Нямдаваа Г. /Ph.D/
Нямжав Б. /М.А/
Дэлгэрзаяа Э. /М.А/

Ховдский филиал Монгольского Государственного университета

Abstract: The development and expansion of the volume of the corpus of Mongolian languages has become an urgent area of linguistics in recent years. In Russia, the corpus of Mongolian, Kalmyk and Buryat languages has been developed on a single platform with the support of the Fundamental Research Program of the Presidium of the Russian Academy of Sciences "Corpus Linguistics". Separate Mongolian language bodies are also being created abroad, while the resources of written languages of the Mongolian peoples prevail. Oral corpus of Mongolian languages, including those collected on the basis of recorded speech of speakers of various dialects, with the necessary transcription and markup, as well as video recording of natural communication between several participants, together with vocal and kinetic annotation of these recordings, are currently becoming extremely relevant due to the rapid "extinction" of the languages of indigenous small peoples of Western Mongolia. The project plans to create an oral (multi-channel) Khalkha-Mongol corps.

Key words: oral corpus, multichannel communication, Oirat dialects, documentation of endangered languages, digital language research

Аннотация

Разработка и расширение объема корпусов монгольских языков становится актуальным направлением языкознания в последние годы. В России к настоящему времени на единой платформе разработаны корпусы монгольского, калмыцкого и бурятского языков при поддержке Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика». За рубежом также создаются отдельные корпуса монгольских языков, при этом преобладают ресурсы письменных языков монгольских народов. Устные корпуса монгольских языков, в том числе собранные на основании записанной речи носителей различных диалектов, с необходимой транскрипцией и разметкой, а также видеозаписью естественной коммуникации между несколькими участниками, в совокупности с вокальной и кинетической аннотацией данных записей, в настоящее время становятся крайне актуальными в связи со стремительным «вымиранием» языков коренных малочисленных народов Западной Монголии. В рамках проекта запланировано создание устного (мультимедийного) корпуса халха-монгольского языка и ойратских диалектов Западной Монголии. Выбор именно мультимедийной коммуникации как объекта изучения связан с тем, что ойраты Западной Монголии в настоящее время не имеют своего письменного языка, используя современный монгольский язык, в то время как процесс «живой» (мультимедийной) коммуникации эфемерен и протекает, не оставляя материальных следов. Кроме этого, цифровая фиксация разных каналов коммуникации полезна для исследователей из разных областей знания. Создание корпуса позволит расширить возможности изучения не только языка и диалектов, но и фольклора, ценностных ориентаций, культурных традиций, обычаев и быта коренных народов Западной Монголии. При создании корпуса в качестве информантов привлекаются представители титульного халха-монгольского этноса, проживающие на территории Западной Монголии, а также ойратских народностей, в том числе представители особой группы коренных малочисленных народов Западной Монголии – хотонов. С фундаментальной точки зрения создание данного цифрового ресурса позволит построить модель мультимедийной коммуникации на материале монгольского языка и ойратских диалектов. Прикладная значимость проекта состоит в документировании идиомов коренных малочисленных народов Западной Монголии, находящихся на грани исчезновения.

Ключевые слова: устный корпус, мультимедийная коммуникация, халха-монгольский язык, ойратские диалекты, документирование исчезающих языков, цифровые исследования языка

До настоящего времени четких критериев, разграничивающих языки, диалекты и говоры Западной Монголии, до сих пор не установлено. В научной литературе зачастую принимается их классификация в зависимости от территории компактного проживания отдельных народностей, исторических оснований, лингвистических особенностей (как правило, лексических или фонетических), также социокультурной среды. Так исследователь Ю. Цэндээ выделяет следующие ареалы компактного проживания ойратов в Западной Монголии: 1) дербеты (аймаки Увс, Баян-Ульгий и Ховд); 2) хотоны (аймак Увс); 3) торгуты (аймак Ховд); 4) баяты (аймак Увс); 5) хошуты (аймак Ховд); 6) захчины (аймак Ховд); 7) урянхайцы (аймаки Баян-Ульгий и Ховд); 8) олёты (аймаки Ховд, Архангай и Говь-Алтай); 9) мянгаты (аймаки Ховд и Увс) (Цэндээ, 2012, 31-54). Для этих ареалов Цэндээ выделяет следующие ойратские диалекты: 1) дербетский (для дербетов и хотонов), 2) баятский, 3) торгутский (для торгутов и хошутов), 4) захчинский, 5) урянхайский, 6) олётский и 7) мянгадский. Также она выделяет группу «халхаизированных» ойратских диалектов, куда, в частности, вошел диалект олётов и хойтов аймаков Архангай, Говь-Алтай и Баянхонгор, а также некоторые другие диалекты. (Цэндээ, 2012, с. 31-60). Некоторые исследователи выделяют хотонский говор дербетского диалекта как особое явление, однако достоверных данных о его существовании на данный момент не существует. Тем не менее, существует ареал их компактного проживания в сомонах Тариалан и Наранбулаг в аймаке Увс. В связи с изложенными тезисами авторы проекта выделяют особую группу ойратских диалектов Западной Монголии и отдельно хотонский говор в качестве специального предмета изучения.

Многие лингвисты признают общность ойратских диалектов Западной Монголии на основе длительных языковых контактов, а также отмечают значительное влияние на них халха-монгольского диалекта и литературного письменного монгольского языка на основе кириллицы (Рассадин и соавт., 2010). При этом влияние ойратской письменности («Тодо-бичиг») оказало лишь локальный и незначительный эффект на устные диалекты Западной Монголии (только на уровне деятельности монастырей, канцелярий и других структур, использующих переводную буддистскую литературу). Кроме того, ойраты Западной Монголии уже долгое время используют монгольскую (кириллическую) письменность, что несомненно наложило и отпечаток на устный вариант их языка, а также на ускорение процесса утраты ойратских диалектов. В связи с этим в лингвистической науке возникала актуальность их документирования и сохранения в цифровом формате. Немаловажную роль в данном направлении могут сыграть создаваемые для разных целей корпуса языков и диалектов.

Корпусные исследования монгольских языков и цифровое документирование диалектов в настоящее время проводятся несколькими группами специалистов. Наиболее представительные российские разработки на материале монгольского языка представлены С. А. Крыловым в корпусе монгольского языка базе данных «Квантитативно-реализационный грамматический словарь современного монгольского языка» (<http://starling.rinet.ru/cgi-bin/main.cgi?root=mongqrs&encoding=utf-rus>), которая представляет полную формальную репрезентацию связей между грамматикой и лексикой монгольского языка в их квантитативном аспекте путём установления квантитативных отношений между единицами языковой системы и дискурсивными единицами, т.е. их речевыми реализациями в текстах. Созданная база данных позволяет составить представление о современном монгольском языке в его количественном аспекте, включая квантитативную модель морфологии, квантитативную модель лексики, квантитативную модель синтаксиса и квантитативную модель стилистики текстов разных жанров. База может быть использована в ходе реализации проекта при анализе записываемых дискурсивных единиц. Тем не менее, создание устных корпусов монгольских языков до сих пор является редкостью, зачастую такие разработки имеют узкую специфику, например, для систем преобразования текста в речь (Batchuluun B., Batchuluun D., & Gombosuren G., 2006; Davaatsagaan & Paliwal, 2008; Sukhbaatar & Munkhtuya, 2006, Ayush & Bayanduren, 2009). Известен лишь один устный диалектный монгольский корпус (MDSCM), в котором представлена вокальная и грамматическая разметка, а также реализована функция автоматического распознавания трех диалектов – халхасского, Цахарского и ойратского диалекта Калмыкии и Синьцзяна (Dawa, et. al., 2006). Корпус был составлен японскими разработчиками для задач машинного перевода, автоматического распознавания речи, а процесс транскрибирования и аннотирования до настоящего времени не завершены. По своей сущности данный корпус не представляет мультимедийную модель коммуникации, а также не включает разнообразные ойратские

диалекты Западной Монголии. Коллекции записей ойратских диалектов создаются также монгольскими исследователями, например, создана коллекция записей Ю. Цэндээ (<http://elar.soas.ac.uk/deposit/0310>). Кроме этого, развиваются корпусные направления исследований на материале других монгольских языков, например, бурятского (Бадмаева, 2009; Цыренов, 2012 и др.), калмыцкого языков (Куканова, 2011).

В рамках документирования ойратских диалектов Западной Монголии также проводятся актуальные исследования, например, в работе Бирталана Агнеса «Лабораторный ойратский: некоторые наблюдения над диалектом алтайских дэрбэтов, записанным в особых условиях» (в журнале «Acta Linguistica Petropolitana. Труды института лингвистических исследований» 2015 г.) обсуждаются вопросы методологические основы создания текстового корпуса диалекта алтайских дэрбэтов, а также достоинства и недостатки предлагаемого подхода. Среди недостатков выделяются следующие: неестественность речи информанта, постоянная коррекция ошибок, влияние литературного монгольского языка у носителей элитарной языковой нормы, «наложение» металингвистических компетенций информанта при порождении устных текстов. При этом автор указывает на перспективность создания более крупного корпуса ойратских диалектов с учетом социолингвистических характеристик информантов, а также на учет сопоставимости в корпусе устной реализации носителями разных диалектов сходных или одного и того же текста (Birtalan, 2012). В рамках проекта впервые запланировано создание устного мультимедийного корпуса монгольского языка и ойратских диалектов Западной Монголии. Выбор именно мультимедийной коммуникации как объекта изучения связан с тем, что ойраты Западной Монголии в настоящее время не имеют своего письменного языка (Рассадин и соавт., 2010) (например, в отличие от ойратов Синьцзяна, по-прежнему использующих «Тодо-бичиг»), используют в письменной коммуникации современный монгольский литературный язык, в то время как процесс «живой» (мультимедийной) коммуникации эфемерен и протекает, не оставляя материальных следов. Кроме этого, разные каналы коммуникации полезны для исследователей смежных областей знания – когнитивной лингвистики, этнолингвистики, этнопсихологии, этнографии. Поэтому создание корпуса позволит расширить возможности изучения не только языка и диалектов, но и психологии и культуры ойратов Западной Монголии. При создании корпуса в качестве информантов привлекаются представители мажоритарной халха-монгольской национальности, проживающие на территории Западной Монголии, а также ойратских народностей, в том числе особая группа коренных малочисленных народов Западной Монголии, предположительно, тюркского происхождения – хотоны.

С фундаментальной точки зрения создание данного цифрового ресурса позволит построить модель мультимедийной коммуникации на материале монгольского языка и ойратских диалектов. Прикладная значимость проекта состоит в документировании языков и диалектов, находящихся на грани исчезновения.

Использование русского языка для перевода устных текстов станет надежной опорой для дальнейших сравнительно-типологических исследований языковых контактов народов Западной Монголии, Алтая, Киргизии, а также для сравнительных исследований в области психолингвистики и когнитивной лингвистики.

Кроме того, проблематика проекта связана с вопросом языковой аттриции (утратой родного языка) среди коренных малочисленных народов, что является одним из закономерных последствий двуязычия и многоязычия (Малышевская и соавт., 2021).

Современный мультимедийный подход к лингвистическому исследованию коммуникации предполагает, что традиционное выделение исключительно вербального кода как предмета научного интереса представляет собой ограниченный взгляд. В реальной коммуникации между людьми не существует принципиальной границы между словами, интонациями, жестами, выражениями лица и т.д. Для участников коммуникации важно, чтобы была достигнута коммуникативная цель. Таким образом, естественный дискурс по своей природе мультимодален. Предлагаемый к разработке устный корпус халха-монгольского языка и ойратских диалектов опирается на теоретические подходы к созданию мультимедийного корпуса русского языка, разработанные в Институте языкознания РАН под руководством А. А. Кибрика, а также опубликованные научные работы данного коллектива (Kibrik & Fedotova, 2018). В проекте используется метод извлечения рассказов (пересказов) на основе специально

снятого для научных целей видеоролика У. Чейфа (в том числе кросскультурных исследований), «The pear stories» («Фильм о грушах»). В этом фильме, не включающем никаких речевых действий персонажей, показана история о взаимодействии ряда лиц, в том числе садовника, собирающего груши, мальчика, крадущего корзину с грушами, и еще нескольких лиц. Фильм представляет собой хорошо продуманную линию событий и часто используется лингвистами как способ получать компактные и сравнимые между собой пересказы. В данном фильме видеоряд подобран так, чтобы респонденты порождали разные рассказы – описания пейзажа, насыщенной сюжетной линии, причинно-следственных отношений, мыслей и эмоций героев и т.д. Теоретические аспекты данного метода достаточно полно описаны в монографии О.В. Федоровой (2014). Известно, что извлеченные пересказы были получены исследователями на нескольких языках, однако материалы устной монгольской речи до сих пор не были задействованы. Важно отметить, что данный метод позволяет приблизиться к пониманию сути мультимодальной коммуникации через ее корпусное моделирование, через призму разных языков и культур, что определяет ценность собираемой коллекции устных рассказов о грушах на халха-монгольском языке и ойратских диалектах в рамках заявляемого проекта. Используются разработанные в Институте языкознания РАН под руководством А.А. Кибрика принципы сбора материала для корпуса и аннотирования данных (<http://multidiscourse.ru/annotation/>).

Монгольский корпус предстоит составить из отдельных коммуникативных эпизодов (так называемых записей) с участием рассказчика (narrator), комментатора (commentator), пересказчика (reteller) и слушателя (listener). Фильм просматривается коммуникантами, которые в дальнейшем выполняют роли рассказчика и комментатора. Рассказчик рассказывает пересказчику содержание фильма в режиме монолога. Затем наступает интерактивный этап, в течение которого комментатор дополняет или уточняет рассказ, а пересказчик задает вопросы обоим собеседникам, посмотревшим фильм. После этого появляется слушатель, и пересказчик пересказывает ему фильм, опять в режиме монолога. В конце слушатель должен письменно записать еще один, вторичный, пересказ фильма. Два последних этапа необходимы для того, чтобы мотивировать всех участников к полноценной и осмысленной коммуникации.

При аннотировании корпуса в рамках реализации проекта планируется реализовать в пилотажном режиме 1-3% основной разметки, включающей аннотирование вокального компонента (вербальный и просодический каналы). Дальнейшее аннотирование корпуса (в том числе дополнительное аннотирование жестов тела, мимики, проксемики, фонетическую транскрипцию, а также референциальную аннотацию) планируется осуществлять совместно с преподавателями Ховдского государственного университета в рамках государственных грантов Монголии в течение нескольких лет, а также студентами данных учебных заведений в рамках курсовых и дипломных проектов, диссертаций. Планируется активное привлечение к участию в проекте, прежде всего, школьных учителей Ховдского, Баян-Ульгийского и Увсунурского аймаков Западной Монголии, а также г. Улан-Батора.

Литература

- Ayush A. & Bayanduren D. Mongolian Speech corpus // PAN localization Project Phase II, 2009
- Batchuluun B., Batchuluun D., and Gombosuren G. (2006) Mongolian text to speech conversation tool. Project Ref. No: 0501A2_L13, Interim report, InfoCon Co., Ltd, Mongolia
- Birtalan Á (2012) Oirad – The Prestige Language: Ethnolinguistic Approach to the Altai Oirad Language and its Dialects // Á. Birtalan (ed.). Oirad and Kalmyk Linguistic Essays [Talentum sorozat 11]. Budapest: ELTE Eötvös Kiadó, 2012. P. 59–74
- Davaatsagaan M. & Paliwal K. Diphone-Based Concatenative Speech Synthesis System for Mongolian // Int'l Multi-Conference of Engineers and Computer Scientists (IMECS) 2008, Vol I, 2008
- Dawa I., Husal, Yue L., Ming Y. Y., Uulang, Cheng B. S., Batsaihan, Arai Y., Mitsunaga M., Isahara H., and Nakamura S. (2006) Multilingual Text–Speech Corpus of Mongolian, Int'l Symposium on Chinese Spoken Language Processing (ISCSLP 2006)
- Kibrik A., Fedorova O. (2018) Empiricheskoe issledovanie mul'tikanal'noy kommunikatsii: russkie rasskazy i razgovory o grushakh. An Empirical Study of Multichannel Communication: Russian Pear Chats

- and Stories. Psychology. Journal of Higher School of Economics, vol. 15, no 2, pp. 191-200
- Sukhbaatar B., and, Munkhtuya D. A New Approach for The Mongolian Text-To-Speech Conversion” // Proc. Int’l Conference on Electronics, Information, and Communication (ICEIC 2006), Vol. 2, pp.303 – 306, 2006;
- Бадмаева Л.Д. Бурятский язык и корпусная лингвистика // Состояние и перспективы развития бурятского языка, мат. форума бурят. яз, Улан-Удэ, 2009
- Крылов С. А. Квантитативно-реализационный грамматический словарь современного монгольского языка. URL: <http://starling.rinet.ru/cgi-bin/main.cgi?root=mongqrs&encoding=utf-rus>
- Куканова В.В. Архитектура метаописания национального корпуса калмыцкого языка // Вестник Калмыцкого института гуманитарных исследований РАН, 2011; http://web-corpora.net/KalmykCorpus/search/?interface_language=ru
- Малышевская А.С., Галло Ф., Бермудез-Маргаретто Б., Штыров Ю.Ю., Читая Т.Д., Петрова А.А., Мичиков А.В. Языковая аттриция: механизмы возникновения, особенности изучения и перспективы дальнейших исследований [Электронный ресурс] // Современная зарубежная психология. 2021. Том 10. № 1. С. 111–124. doi:10.17759/jmfp.2021100111
- Рассадин В.И. и др. Историческая связь калмыцкого языка с языком ойратов Монголии. Элиста, 2010
- Федорова О.В. Экспериментальный анализ дискурса, М., 2014
- Цыренов Б.Д. О практической и теоретической лексикографии бурятского языка (к истории вопроса) // Вестник Калмыцкого института гуманитарных исследований РАН, 2012. С. 24-30
- Цэндээ Ю. Ойрад аялгууны хэлзүй: авиа зүй, үг зүй, өгүүлбэр зүй, Sojombo, 2012, с. 31-60.



“ХЭЛ, СОЁЛ, СОЁЛ ХООРОНДЫН ХАРИЛЦАА”

Олон улсын эрдэм шинжилгээний VIII хурлын эмхэтгэл

“LANGUAGE, CULTURE, INTERCULTURAL COMMUNICATION”

8TH INTERNATIONAL SCIENTIFIC CONFERENCE

Улаанбаатар
2022

ННА 80
ДДА 400
Х-892

Зохион байгуулагч: МУИС-ийн Ховд аймаг дахь салбар
сургууль
МУИС-ийн Ховд аймаг дахь салбар
сургуулийн Күнзийн институт
Гадаад хэл, соёл судлалын тэнхим

Редакцийн зөвлөл: Г.Амарзаяа
МУИС-ийн Ховд аймаг дахь салбар
сургуулийн ЭШХАА-ны дарга, доктор,
дэд профессор
М.Алтансүвд
МУИС-ийн Ховд аймаг дахь салбар
сургуулийн багш, доктор
В.Янжиндулам
МУИС-ийн Ховд аймаг дахь салбар
сургуулийн багш, доктор
Ж.Ганцоож
МУИС-ийн Ховд аймаг дахь салбар
сургуулийн дэргэдэх Күнзийн
институтын Монгол талын захирал,
магистр

Хавтасны дизайнер: У.Билэгт-Од

Цаасны хэмжээ: 70x100/16

Хэвлэлийн хуудас: 40.5 хх

Хэвлэлийн "Соёмбо Принтинг" ХХК-д хэвлэв.

ISBN 978-9919-9788-1-5

© Зохиогчийн зөвшөөрөлгүй хэвлэн олшруулахыг хориглоно.