Абаева Юлия Догоржаповна

ФОНЕТИЧЕСКАЯ БАЗА ДАННЫХ БУРЯТСКИХ ДИАЛЕКТОВ: ПРИНЦИПЫ СОСТАВЛЕНИЯ И РАЗМЕТКИ

Статья посвящена научно-исследовательскому проекту по созданию фонетической базы данных бурятских диалектов. В статье рассматриваются основные этапы работы, принципы составления базы данных, диалектный материал, который войдет в корпус, приводятся основные параметры разметки. Особенность разрабатываемого проекта в том, что современное состояние звукового облика бурятских диалектов впервые представлено в формате базы данных.

Адрес статьи: www.gramota.net/materials/2/2015/1-1/1.html

Источник

Филологические науки. Вопросы теории и практики

Тамбов: Грамота, 2015. № 1 (43): в 2-х ч. Ч. I. С. 13-16. ISSN 1997-2911.

Адрес журнала: www.gramota.net/editions/2.html

Содержание данного номера журнала: www.gramota.net/materials/2/2015/1-1/

УДК 811.512.31
Филологические науки

Статья посвящена научно-исследовательскому проекту по созданию фонетической базы данных бурятских диалектов. В статье рассматриваются основные этапы работы, принципы составления базы данных, диалектный материал, который войдет в корпус, приводятся основные параметры разметки. Особенность разрабатываемого проекта в том, что современное состояние звукового облика бурятских диалектов впервые представлено в формате базы данных.

Ключевые слова и фразы: база данных звуковых файлов; речевой корпус; диалекты бурятского языка; сегментация звуковых файлов; компьютерная обработка звуковых файлов; параметры разметки.

Абаева Юлия Догоржаповна
Институт монголоведения, буддологии и тибетологии Сибирского отделения Российской академии наук
julaba@yandex.ru

ФОНЕТИЧЕСКАЯ БАЗА ДАННЫХ БУРЯТСКИХ ДИАЛЕКТОВ: ПРИНЦИПЫ СОСТАВЛЕНИЯ И РАЗМЕТКИ©

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект № 12-06-98014р_сибирь_а.

В настоящее время задача создания больших, разнообразных и информативных (многоуровневых) речевых баз данных становится все более актуальной. Электронная база данных представляется оптимальным способом хранения, систематизации и управления большим объемом лингвистической информации. Базы данных, содержащие большие массивы текстов, подвергнутые лингвистической разметке и снабженные специализированной системой управления, иначе называемые корпусами, получили большое распространение в рамках корпусной лингвистики: к примеру, это Британский национальный корпус (British National Corpus), Национальный корпус русского языка [9], Корпус бурятского языка [1; 4]. Корпусы могут отличаться по своим целям, способу представления материала, это могут быть корпусы устной речи [11; 13], корпусы диалектных текстов [8; 10], исторические и современные (Национальные корпусы испанского языка *CORDE* и *CREA*) [6], специализированные корпусы по фонетике, например корпус, содержащий информацию об интонации диалектов английского языка, «Интонационная вариативность в английском» (The IViE Corpus) [14] и т.д. Создание речевых баз данных актуально как для широко распространенных активно используемых языков, так и для миноритарных языков и языков, находящихся на грани исчезновения [5; 7; 10].

П. А. Скрелин и П. П. Щербаков, говоря о требованиях к современной фонетической базе данных, указывают, что звуковая база данных должна содержать звуковой материал, «представляющий максимальную вариативность реализации языковых единиц в различных условиях речевой деятельности человека, сегментную информацию и подробное описание представленного звукового материала» [12, с. 62]. В нашем исследовании мы руководствовались этими требованиями. На первом этапе работы над созданием фонетической базы данных территориальных вариантов бурятского языка, были определены цели, на решение которых направлено создание информационного ресурса, а также решались следующие задачи: определено количество диалектов, уточнено распределение говоров по диалектным группам; учтены фонетические особенности диалектов; разработан корпус, отражающий выявленные особенности; осуществлен подбор дикторов, запись материала; проведена систематизация, редактирование и структурирование данных.

Цель создания фонетической базы данных бурятского языка с учетом его территориальной вариативности – сохранение живой разговорной бурятской речи во всем многообразии ее произносительных вариантов. Звуковой диалектный материал позволит зафиксировать современное состояние диалектов, их уникальные речевые особенности.

Бурятский язык многодиалектный. Буряты проживают достаточно разрозненно на обширных территориях в пределах трех государств: России (Республика Бурятия, Иркутская область, Забайкальский край), Китая, Монголии, что обусловило различия на всех языковых уровнях, в том числе и на фонетическом (как сегментном, так и суперсегментном). В настоящее время эти особенности стремятся к нивелированию под влиянием различных факторов. С одной стороны, это влияние родственных доминирующих языков: литературный бурятский язык в Бурятии, халха-монгольский в Монголии, близкородственные монгольские языки, в том числе литературный восточно-монгольский язык, в Китае. С другой стороны, происходит общее сужение сфер употребления бурятского языка, что обусловлено различными экстралингвистическими факторами.

На данном этапе материал базы данных ограничен диалектами бурят, проживающих на территории Российской Федерации в так называемом Байкальском регионе (к западу, востоку и югу от оз. Байкал). Соответственно, наиболее уместен в данном случае территориальный принцип деления диалектов на западные, восточные и южные. Такое деление принято в большинстве диалектологических работ по бурятскому языку и различается лишь по составу входящих в них говоров. В вопросе о принадлежности того или иного говора к территориальной группе мы придерживались классификации, предложенной И. Д. Бураевым [3]. Он совмещает территориальный принцип деления с таким немаловажным фактором, как принадлежность к той или иной родо-племенной группе бурят, а также учитывает конкретные языковые факты.

Так, к восточной группе им относится говор хоринских бурят, в том числе агинских, а также иволгинских и северноселенгинских бурят, которые не являются хоринцами по происхождению, но имеют большое количество сходных черт в языке, обусловленных длительным контактированием.

К западной группе помимо эхирит-булагатских, боханских, ольхонских и качугских бурят, проживающих в Прибайкалье, отнесены говоры баргузинских и байкало-кударинских бурят, которые территориально расположены к востоку от оз. Байкал.

Южная группа включает цонгольский и сартульский говоры, а также И. Д. Бураевым сюда отнесена «цакающая часть ононских хамниган» [Там же, с. 24].

Помимо этих больших групп в описываемой классификации выделяется достаточно крупная группа хонгодорских родов, говорящих на аларо-тункинском наречии, а также два небольших самостоятельных говора: крайний западный (говор нижнеудинских бурят) и крайний восточный (говор ононских хамниган).

Анализ фонемного состава диалектов бурятского языка позволил уточнить наиболее заметные их отличия. В результате был составлен список всех фонем говоров и диалектов бурятского языка. Разработана система записи звуков диалектов бурятского языка при помощи системы *SAMPA* (Speech Assesment Methods Phonetic Alphabet). Фонетический алфавит *SAMPA* позволяет записать фонетические особенности языка при помощи знаков *Word*, не используя специальные фонетические знаки, которые присутствуют не во всех компьютерах и тем самым создают сложности при работе с файлами на разных компьютерах.

В дальнейшем был составлен речевой корпус, или программа, которую предполагалось озвучить с помощью дикторов. Речевой корпус состоит из разноуровневых речевых фрагментов и призван выявить наиболее яркие характерные черты того или иного диалекта. Основные требования к речевому корпусу были следующими.

Слова. Моносиллабы должны покрывать все возможные типы слогов (V, VC, VCC, CV, CVC, CVCC), включать все гласные и согласные и их комбинации. Дисиллабы и полисиллабы представляют собой слова-основы и их производные с наибольшим количеством фонетических вариантов (все типы слогов и их сочетаний, сингармонизм, коартикуляция). Фонемы по возможности должны быть представлены во всех позициях в слове.

Предложения. Образцы разговорной речи на бытовые темы, включающие все возможные интонационные типы.

Тексты. Повествования монологического характера на повседневные бытовые темы, исторические события, мифы и предания.

При разработке речевого материала для записи и последующего анализа использовались методы и подходы корпусной лингвистики. Речевой корпус записан в виде файлов *Word* и *Excel*. Таблица для корпуса слов содержит следующие графы: порядковый номер, орфографическая запись, фонетическая запись в системе МФА и *SAMPA*, перевод на русский язык, количество слогов, слоговой состав, часть речи. Корпус предложений, представленный в виде файла *Excel*, имеет графы: порядковый номер, орфографическая запись, фонетическая запись в системе МФА и *SAMPA*, перевод на русский язык, коммуникативный тип, количество слов.

Материал записывался во время командировок в места проживания носителей говора: Иркутская область (Нукутский и Эхирит-Булагатский районы), Республика Бурятия (Джидинский район, с. Верхний Бургултай, Нижний Бургултай, Цагатуй), Забайкальский край (пос. Агинское). Аудиозапись осуществлялась при помощи цифровых диктофонов с использованием высокочувствительных микрофонов. Параметры записи 22 *kHZ*, 16 *bit*. Для каждого территориального варианта было записано по 4 диктора: 2 мужчины и 2 женщины.

Далее записанный материал был нарезан на отдельные звуковые файлы, содержащие одно слово, одно предложение или один текст, и каждому файлу был присвоен индекс. Слова индексировались следующим образом: *LitM1A001*, где *Lit* указывает на диалектную принадлежность слова (в данном случае – к литературному варианту), *M1/F1* – пол диктора с его порядковым номером, буквы *A-D* указывают на количество слогов (A – 1, B – 2, C – 3, D – 4 и более слогов). Для файлов, содержащих предложения, индексы, помимо диалектной принадлежности, номера диктора и порядкового номера, указывают на коммуникативный тип, например: *SarF2Nar001* (сартульский говор, диктор женского пола № 2, повествование). Далее была создана База данных индексов в виде файла *Excel*, которая включает следующие графы для слов: № словоформы, индекс, перевод на русский язык, транскрипция МФА и *SAMPA*, количество слогов, типы слогов. Для предложений: № словоформы, индекс, перевод на русский язык, транскрипция МФА и *SAMPA*, коммуникативный тип, количество слов.

Компьютерный анализ записанного материала проводится при помощи программы анализа и синтеза речи *PRAAT*, разработанной учеными из Нидерландов и предназначенной главным образом для фонетических исследований. Эта программа позволяет получать спектрограмму, графики ЧОТ и интенсивности во временной зависимости, с возможностью редактирования звуковых сегментов и аннотирования. Звуковой файл подвергается многоуровневой сегментации: 1 уровень – слово, 2 уровень – слоги, 3 уровень – фонемы. Сегментация выполняется вручную, поэтому это достаточно трудоемкий и длительный процесс, требующий определенных навыков.

Следующим этапом работы над базой данных было создание системы таблиц, поля которых отражают результаты разметки описываемых объектов. Разметка, или аннотация, – это важный этап при формировании базы данных. Именно наличие этой дополнительной информации (о диалекте, дикторе, сегментном и суперсегментном составе и т.д.) превращает совокупность речевых фрагментов в базу данных [2]. От качества разметки зависит возможность выполнения поисковых запросов, что и определяет теоретическую и практическую ценность выполненной работы.

В качестве основы программного обеспечения выбрана реляционная база данных *MS Access* для *Windows*. Этот формат позволяет редактировать базу и, при необходимости, вводить в нее новые языковые данные. Кроме того, эта программа является общедоступной, имеет достаточно простой интуитивно понятный интерфейс, что немаловажно для неподготовленного пользователя.

Таблицы, в которых отражены результаты разметки записанного материала, можно разделить на две группы. Прежде всего, это информация общего типа о дикторах (их диалектной принадлежности, составе записанного от них материала) и об исходных звукозаписях (длительность звучания, содержание, качество и т.д.). Другая группа таблиц может быть охарактеризована как научно-исследовательская, то есть содержит результаты компьютерной обработки звуковых файлов.

Таблица 1.

Общая информация

Данные об информантах	**Звуковые файлы: информация об исходных звуковых файлах, полученных от информантов**
• Порядковый номер • Фамилия, имя, отчество, или псевдоним • Код информанта (M1, F1) • Пол • Возраст (год рождения, точный или приблизительный) • Место рождения • Социальное происхождение • Образование • Профессия • Место постоянного проживания • Родовая принадлежность • Язык и диалект постоянного общения • Место записи • Год записи • Комментарии (информация относительно типа личности, речевых особенностей и т.п.) • Информация о звуковых файлах, полученных от информанта	• Код информанта • Номер диктофонной записи (полученный при записи файла на диктофон) • Общее время звучания • Общая информация о содержании (запись программы, рассказ о себе, диалоги на бытовые темы и т.п.) • Комментарии к файлу • Наличие компьютерной обработки

Таблица 2.

Результаты компьютерной обработки звуковых файлов

Результаты сегментации словоформ на уровни	**Ударение: результаты компьютерного анализа**
Слово: порядковый номер № словоформы в базе индексов транскрипция *IPA* и *SAMPA* Слоги: количество слогов в слове транскрипция *IPA, SAMPA* тип слога позиция слога в слове (1, 2… *fin*) Фонемы: транскрипция *IPA, SAMPA* позиция в слове (порядковый номер) позиция в слоге (порядковый номер) дополнительные характеристики (*schwa*, оглушение, выпадение, придыхание и т.п.) Фонация: тип фонации (модальный, скрипучий, пониженный, придыхательный, фальцет, шепотный и т.д.)	• № файла в системе индексов • Орфографическая запись • Транскрипция *IPA, SAMPA* • Общая длительность звучания • Количество слогов • Информация о ЧОТ слогов • Информация о длительности слогов • Информация об интенсивности слогов
Интонация: результаты компьютерного анализа	**Тексты: описание темы повествования**
• № файла в системе индексов • Орфографическая запись • Транскрипция *IPA, SAMPA* • Общая длительность звучания • Тип высказывания (повествование, общий вопрос, частный вопрос, переспрос, побуждение) • Количество слов • Информация о ЧОТ слогов • Информация о длительности слогов • Информация об интенсивности слогов	• Имя файла • Тематика текста (примерное содержание повествования, например Семья, Труд, Обряды, Праздники, Погода, Природа и т.п.) • Жанровая принадлежность (биографический рассказ, описание, интервью, рассуждение, фольклорное произведение и т.п.) • Время звучания • Наличие/отсутствие расшифровки • Комментарии В дальнейшем планируется снабдить каждый текст расшифровкой орфографической и фонетической

Работа по созданию фонетической базы данных бурятских диалектов позволит расширить спектр исследований в области бурятской сегментной и суперсегментной фонетики, диалектологии, зафиксировать современное состояние бурятских диалектов, проследить основные процессы, происходящие в речи современных бурят. Эта работа может войти в качестве диалектного звукового подкорпуса в Корпус бурятского языка.

Список литературы

1. **Бадмаева Л. Д.** Бурятский языковой корпус: создание, проблемы и перспективы // Вестник Бурятского научного центра Сибирского отделения Российской академии наук. 2013. № 2 (10). С. 118-122.
2. **Богданов Д. С., Кривнова О. Ф., Подрабинович А. Я., Фарсобина В. В.** База речевых фрагментов русского языка –ISABASE" [Электронный ресурс]. URL: http://kk.convdocs.org/docs/index-101775.html (дата обращения: 22.08.2014).
3. **Бураев И. Д.** Основные этапы исследования бурятских диалектов и их классификация // Развитие и взаимодействие диалектов Прибайкалья. Улан-Удэ, 1988. С. 3-25.
4. **Бурятский корпус** [Электронный ресурс]. URL: http://web-corpora.net/BuryatCorpus/search/index.php?interface_language=ru (дата обращения: 25.08.2014).
5. **Долозова О. Н.** Создание и лингвистическая разметка звуковой словарно-грамматической базы данных по ительменскому языку [Электронный ресурс]. URL: http://www.dialog-21.ru/digests/dialog2010/materials/html/18.htm (дата обращения: 22.08.2014).
6. **Жолобова А. О.** Национальный корпус испанского языка: *CORDE* и *CREA* // Филологические науки. Вопросы теории и практики. Тамбов: Грамота, 2014. № 9. Ч. 1. С. 56-58.
7. **Казакевич О. А.** Мультимедийная база данных исчезающего языка [Электронный ресурс]. URL: http://www.dialog-21.ru/Archive/2001/volume1/1_17.htm (дата обращения: 22.08.2014).
8. **Летучий А. Б.** Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003-2005: результаты и перспективы. М., 2005. С. 215-232.
9. **Национальный корпус русского языка** [Электронный ресурс]. URL: http://ruscorpora.ru/ (дата обращения: 22.08.2014).
10. **Некрасова Г. А.** Электронный диалектный корпус как ресурс сохранения и изучения коми диалектов // Финно-угорский мир. 2010. № 1. С. 13-16.
11. **Рассказы о сновидениях и другие корпуса звучащей речи** [Электронный ресурс]. URL: http://spokencorpora.ru/ (дата обращения: 22.08.2014).
12. **Скрелин П. А., Щербаков П. П.** Требования к современной фонетической базе данных для фундаментальных и прикладных исследований // Технологии информационного общества – Интернет и современное общество: труды VI Всероссийской объединенной конференции (Санкт-Петербург, 3-6 ноября 2003 г.). СПб.: Изд-во филологического фак-та СПбГУ, 2003. С. 62-63.
13. **Степанова С. Б., Асиновский А. С., Богданова Н. В., Русакова М. В., Шерстинова Т. Ю.** Звуковой корпус русского языка повседневного общения «один речевой день»: концепция и состояние формирования [Электронный ресурс]. URL: http://www.dialog-21.ru/digests/dialog2008/materials/html/76.htm (дата обращения: 22.08.2014).
14. **The IViE Corpus** [Электронный ресурс]. URL: http://www.phon.ox.ac.uk/files/apps/IViE/ (дата обращения: 22.08.2014).

PHONETIC DATABASE OF THE BURIAT DIALECTS: PRINCIPLES OF FORMATION AND MARKING

Abaeva Yuliya Dogorzhapovna
Institute of Mongolian, Buddhist and Tibetan Studies of Siberian Branch of the Russian Academy of Sciences
julaba@yandex.ru

The article is devoted to the research project on the creation of a phonetic database of the Buryat dialects. The article considers the main stages of the work, the principles of database compiling, and the dialect material that forms the corpus, and provides the basic marking parameters. The project under development has the following distinctive feature: the current state of the sound character of the Buryat dialects is for the first time presented in the format of a database.

Key words and phrases: database of sound files; speech corpus; dialects of the Buryat language; segmentation of sound files; computer processing of sound files; marking parameters.

УДК 81
Филологические науки

В статье рассматривается понятие языковой интерференции. Выявляются ситуации, в которых наблюдается явление интерференции. Анализируются различные уровни (фонологический, грамматический, лексический и орфографический), на которых может рассматриваться интерференция, а также дается характеристика видам интерференции. Приводятся примеры для каждого уровня интерференции. Анализируется роль языковой интерференции при овладении иностранным языком.

Ключевые слова и фразы: интерференция; языковой контакт; немецкий язык; иноязычный коммуникант; речевое общение.

Абрамова Наталья Викторовна, к. пед. н., доцент
Саратовская государственная юридическая академия
nataklenin@mail.ru

ФУНКЦИОНАЛЬНО-СТРУКТУРНЫЕ ОСОБЕННОСТИ ЯВЛЕНИЯ ИНТЕРФЕРЕНЦИИ В УСЛОВИЯХ ЯЗЫКОВОГО КОНТАКТА (НА ПРИМЕРЕ НЕМЕЦКОГО ЯЗЫКА)©

Проблема языкового контакта является актуальной для современного языкознания. Язык – это главное средство, с помощью которого люди контактируют друг с другом, выражают свои чувства. При этом родной язык