ЯЗЫКОЗНАНИЕ И ЛИТЕРАТУРОВЕДЕНИЕ

УДК 004.658:801.3=512.141
DOI: 10.31040/2222-8349-2020-0-2-90-97

РАЗРАБОТКА АУДИОКОРПУСА ВОСТОЧНОГО ДИАЛЕКТА БАШКИРСКОГО ЯЗЫКА: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ

© Л.А. Бускунбаева, З.А. Сиразитдинов

Рассматриваются приципы создания корпуса аудиоматериалов восточного диалекта башкирского языка. Ставится задача репрезентативного сбора полевых материалов с учетом возрастной, гендерной принадлежности, уровня образования, языка общения, национальной идентичности информантов и разнообразия тем для беседы.

В корпусе предполагается представление аудиозаписи, расшифровки звуковых файлов в виде транскрибированных текстов, их литературный вариант и русский перевод диалектного текста.

Авторами предлагается транскрипция в «полуорфографической записи», широко распространенной в речевых корпусах. Разработанная система графической передачи диалектных особенностей основана на использовании специальных знаков и буквосочетаний. Членение тематического монолога информанта на фразы и синтагмы, транскрибирование и синхронизация транскрипта с аудиофайлом осуществляется в программе аннотирования ELAN. Просодические характеристики речи, такие как ударение, интонация, тон, на данном этапе транскрибирования диалектного материала не учитываются.

Ключевые слова: башкирский язык, тюркские языки, лингвистический корпус, корпусная лингвистика, диалектология, экспедиционный материал.

Введение. Одним из приоритетных направлений современного языкознания является корпусная лингвистика, объектом исследования которой являются речевые материалы, реализованные как в виде письменных текстов, так и устных (фонетических) массивов данных, снабженные определенной системой разметок. В количественном соотношении письменные корпусы намного превосходят устные, но в то же время темп составления последних с каждым годом растет, что определяется все более возрастающим интересом исследователей к разговорному тексту, анализу особенностей функционирования языковых единиц в речевом потоке.

Сформированные идеи и подходы корпусной лингвистики начали активно применяться и в диалектологии, где до недавнего времени основной источниковой базой оставался материал, собранный по специальным вопросникам и анкетам и представленный в картотеках, словарях, атласах, монографиях и статьях языковедов. Современная же лингвистика выдвигает требование статистической объективности данных, которое реализуемо при корпусном подходе.

Сегодня часть существующих диалектных корпусных материалов включена в состав национальных корпусов, часть же функционирует самостоятельно. В России известны Саратовский диалектологический корпус, Мультимедийный корпус диалектных текстов Устьянского района Архангельской области, диалектные подкорпусы в составе Национального корпуса русского языка, Национального корпуса калмыцкого языка, ижемского говора коми языка и др.

Следует отдельно отметить уникальный проект, разрабатываемый учеными Института языкознания РАН и Томского государственного университета в сотрудничестве с Институтом системного программирования РАН, который представляет собой информационную систему, содержащую обширную информацию по уральским и алтайским языкам на базе экспедиционных аудиозаписей: корпусы, мультимедийные словари и пр. Аудиозаписи транскрибированы, глоссированы и переведены на русский язык [http://lingvodoc.ispras.ru].

БУСКУНБАЕВА Лилия Айсовна – к.филол.н., Ордена Знак Почета Институт истории, языка и литературы УФИЦ РАН, e-mail: buskl@yandex.ru
СИРАЗИТДИНОВ Зиннур Амирович – к.филол.н., Ордена Знак Почета Институт истории, языка и литературы УФИЦ РАН, e-mail: sazin11@mail.ru

Разработка корпусов диалектных текстов сопряжена с целым рядом сложностей, к которым относятся системные языковые отличия от литературного языка; исключительно устный характер диалектной коммуникации, вариативность на всех уровнях, затрудняющая идентификацию единиц в корпусе; отсутствие единообразия при фиксации диалектной речи и различные способы организации информации [1]. Ввиду специфичности диалектного материала процесс подготовки данных текстов и включение их в корпус более трудоемкий и занимает больше времени и усилий, поскольку включает в себя запись информанта, расшифровку аудиофайла, транскрибирование, ввод в базу данных и разработку экстралингвистических и лингвистических разметок. Как отмечают разработчики Саратовского диалектологического корпуса, «это не механическая, а исследовательская работа, серьезный и очень ответственный авторский труд» [2, с. 363].

Современное состояние изученности восточного диалекта. Ареалом распространения восточного диалекта башкирского языка являются Абзелиловский, Баймакский, Белокатайский, Белорецкий, Бурзянский, Кигинский, Мечетлинский, Салаватский, Учалинский районы Республики Башкортостан, а также прилегающие к ним отдельные районы Челябинской, Курганской и Свердловской областей.

Восточный диалект башкирского языка не раз становился объектом исследования диалектологов. Говоры диалекта изучены методами лингвистической географии, описаны монографически, изданы многочисленные научные исследования, составлены словари.

Монография «Восточный диалект башкирского языка» в сравнительно-историческом освещении известного башкирского диалектолога Н.Х. Максютовой [3] представляет собой первый опыт систематического описания фонетических, лексических и морфологических особенностей рассматриваемого диалекта. Наряду с проблемами становления и развития диалекта в научной работе выделяются пять говоров, описывается современное состояние его говоров, определяются границы их распространения. Однако в работе вне поля исследования остается язык несколько обособленных деревень Учалинского района, который диалектологом Р.З. Шакуровым выделяется в отдельный говор восточного диалекта [4, с. 42].

Восточный диалект подвергался и лексикографическому описанию. Изданный в 1967 г. первый том «Словаря башкирских говоров» [5] посвящен восточному диалекту и включает в себя около 8000 диалектизмов, охватывающих все говоры данного диалекта. «Диалектологический словарь башкирского языка» [6] содержит лексические единицы данного диалекта.

Изданный в 2005 г. «Диалектологический атлас башкирского языка» включает в себя уникальный материал, собранный диалектологами ИИЯЛ УНЦ РАН в 1973–1983 гг. и показывающий территориальные распределения фонетических, лексических и грамматических явлений по 250 опорным пунктам Республики Башкортостан и сопредельных областей [7]. Данный атлас также содержит значительный материал по восточному диалекту.

Часть текстов восточного диалекта была опубликована в виде транскрипции в сборнике [8].

Лексикографические диалектные материалы, диалектологический атлас башкирского языка и транскрибированные тексты из образцов речи представлены в виде диалектологической базы данных в Машинном фонде башкирского языка [mfbl2.ru].

Все комплексные исследования по данному диалекту (по другим диалектам ситуация схожая) проводились до 80-х гг. XX столетия. К сожалению, с 80-х гг. XX в. и по настоящее время диалекты остаются в стороне от фронтальных полевых исследований. Количество экспедиций с единой программой сбора полевого материала резко уменьшилось, большинство материалов прежних экспедиций не было оцифровано, значительная часть их утрачена. Ни корпуса по материалам восточного диалекта, ни самих оцифрованных материалов для создания корпуса не существует.

Данный проект призван начать сбор и обработку обширного материала с охватом широкого круга информантов по поло-возрастным и другим социальным группам с перспективой создания представительного корпуса, включающего в себя аудиоматериалы и их транскрипции по говорам восточного диалекта башкирского языка.

Принципы сбора полевых материалов и разработка метаразметки. Создание репрезентативного диалектного корпуса требует максимального охвата всех говоров и содержательной

металингвистической разметки представленного текста с учетом возрастной, гендерной принадлежности, уровня образования, языка общения, национальности информантов, разнообразия тем для беседы и т.д.

1) Сбор диалектного материала производится по всем 6 говорам восточного диалекта башкирского языка: айский, сальютский, аргаяшский, миасский, кизильский и учалинский.

2) По каждому населенному пункту предполагается осуществить запись минимум 12 информантов: 2 записи по гендерному признаку, 6 записей по возрастным группам. Выделяются следующие возрастные группы:

– дошкольный и начальный класс (до 11 лет);

– средний школьный (от 11–15 лет);

– старший и студенческий возраст (16–25 лет);

– средний возраст (25–45 лет);

– старший возраст (45–65 лет);

– пожилой возраст (от 65 лет).

Учитывается образование информанта: начальное, среднее, высшее.

3) Аудиозапись сопровождается информацией, которая составляет экстралингвистическую разметку аудиофайла:

– гендер: мужской, женский;

– образование: начальное, среднее (среднее школьное или суз), высшее;

– возраст;

– язык обучения: башкирский, русский, татарский, чувашский;

– язык общения в семье: башкирский, русский, татарский;

– национальность информанта;

– имя, отчество, фамилия;

– место проживания;

– место последнего долгого проживания до переезда в данное место (в случае переезда);

– время проживания до последнего места проживания (в случае переезда);

– время записи.

4) Выделяется тип общения: монолог, диалог, полилог.

5) Для записи информантов определены следующие 15 тем:

– свадьба, свадебные обычаи;

– блюда (повседневные и праздничные);

– домашние животные (какие держат и как содержат);

– система родства (дети и близкие родственники);

– приусадебное хозяйство (огород, сад);

– дом, постройки (когда и кем построен, какая крыша, рамы);

– топонимия в окрестностях поселения;

– история села, школы, рода;

– повседневная жизнь (работа, школа);

– времена года, погода;

– малые формы фольклора (частушки, пословицы, поговорки, сказки);

– поездка в райцентр (по каким делам, каким транспортом пользуются);

– игры детей;

– друзья (кто они, где они живут);

– животный мир около поселения (какие птицы и звери обитают).

6) Запись осуществляется на цифровой диктофон в несжатом формате PCM (.WAV), при отсутствии посторонних звуков (16бит/22kHz – 16бит/48kHz). Первичная обработка аудиозаписей (очистка от посторонних шумов и длительных пауз) и паспортизация производятся в программе Sound Forge.

7) Паспортизация файлов и экстралингвистическая разметка осуществляются в базе данных Access, частично включаются и в имена аудиофайлов:

1. v – восточный диалект, aj – айский говор, ar – аргаяшский говор, sl – сальютский говор, mi – миасский говор, kz – кизильский говор; uc – учалинский говор.

2. m – мужчина, w – женщина.

3. Образование: n – начальное образование, s – среднее образование, v – высшее образование.

4. Номер возрастной группы: 1 – дошкольный и начальный класс, 2 – средний школьный, 3 – старший и студенческий возраст, 4 – средний возраст, 5 – старший возраст, 6 – пожилой возраст.

5. t1 – t15 – темы;

Например, аудиофайл vuchmst106d023 означает принадлежность звукового файла восточному диалекту, учалинскому говору, в котором осуществлена запись на тему «времена года, погода» мужчиной пожилого возраста со средним образованием.

Дополнительные коды информанта: от d001 до d999. Этот код является именем текстового файла, в котором указываются имя, фамилия, отчество, язык обучения, язык общения в семье, национальность информанта, место последнего долгого проживания до переезда в данное место (в случае переезда), время проживания до последнего места проживания

(в случае переезда), время проживания в данном месте, дата осуществления записи и др.

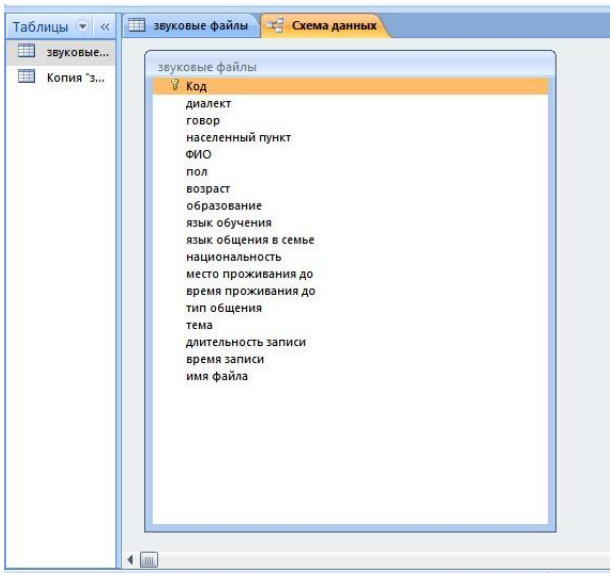

Рис. 1. Схема данных паспортизации файлов и экстралингвистической разметки в базе Access

Принципы транскрибирования и лингвистической разметки. Членение аудиофайла на коммуникативные эпизоды (фразы), транскрибирование и синхронизация транскрипта с аудиофайлом осуществляются в программе аннотирования ELAN. Работа осуществляется в трех уровнях: транскрибирование аудиофайла, представление в литературной форме и перевод на русский язык. Транскрипция выполняется в «полуорфографической записи», широко распространенной транскрипции в речевых корпусных разработках [9]. Эта транскрипция близка к фонематической, максимально приближена к орфографии современного башкирского языка. Была разработана единая система графической передачи диалектных особенностей с использованием специальных знаков и буквосочетаний для обозначения тех звуков, которые невозможно передать при помощи башкирской орфографии.

Гласные фонемы:

Наряду с фонемами ә, о, ө, у, ү, и, ы, э (е) выделены следующие гласные:

а* – гортанный звук (ҡара – ҡа*ра);

а^ – огубленный звук (алма – а^лма).

Согласные фонемы:

Наряду с фонемами б, й, ж, з, ҙ, д, к, ҡ, ғ, г, л, м, н, ң, п, р, с, ҫ, т, ф, х, ш, һ выделены следующие согласные:

v – билабиальная шумная согласная в (ваҡыт – vаҡыт, вәғәҙә – vәғәҙә);

б* – билабиальный вариант согласной [б] (арба – арб*а);

ҡ* – гортанный вариант согласной [ҡ] (ҡалын – ҡ*алын);

ф*– билабиальная согласная, близкая к [ф] (афын – аф*ын);

ҫ* – твердая верхнезубная согласная;

у^ – огубленный вариант гласной [у] (юҡ – йу^ҡ);

ү^ – огубленный вариант гласной [ү] (күп – кү^п);

и* – гласная, близкая к [е] (кил – ки*л);

йа – вместо буквы [я] (яҡ – йаҡ);

йу – вместо буквы [ю] (юл – йул);

йе – вместо буквы [е] в начале слова (емеш – йемеш).

Для различения аллофонов использованы знаки (*, ^) с учетом их ввода в ELAN.

Надо отметить, что вышеуказанная транскрипция применяется только при транскрибировании башкирских слов. Вкрапления из русского языка представлены в оригинальной орфографической записи, если они под влиянием башкирского языка не подвергались видоизменениям: *ана шул (э-э) нейҙәрҙә беҙҙең (э-э) буш ваҡытында / теге # пасиф # ваҡыты бөттө нисә / китереб беҙҙең ҡапҡа төб*өнә өйөб ҡуйа торғандар ийе инде // ‘после посева / когда не было в них необходимости / приносили к нам и складывали возле ворот //’*. Октоторпом (#) в транскрибированных текстах отмечается переключение языковых кодов (с башкирского на русский).

Единицей описания в корпусе являются не только слова, но и знаки сегментирования, символы, обозначающие паралингвистические элементы речи, такие как смех, кашель, вздохи, стоны, причмокивание, плач и другие хезитационные явления, сопровождающие живую речь.

При разработке аннотирования на уровне высказываний в орфографической записи был применен опыт создателей диалектологических и мультимедийных корпусов русского и калмыцкого языков [10–12].

Особое внимание уделяется и непреднамеренным остановкам информационного потока в процессе коммуникации, которые обусловлены целым рядом факторов как индивидуальных, психологических, так и физиологических. Как заметил американский паузолог О'Коннор, «паузы могут сказать о человеке не меньше,

чем слова, и в разговоре тратится 40–50% времени на них» [Цит. по: 13]. Поэтому разработчиками было решено зафиксировать паузы хезитации при транскрибировании диалектных текстов.

В большинстве случаев такие паузы возникают в случае наличия у говорящего определенных сложностей «в планировании текущего высказывания – чем больше этих сложностей, тем более вероятно появление паузы хезитации и тем больше ее продолжительность» [13].

В потоке речи информанта различают короткую паузу ( ) и длительную (..). Паузы хезитации могут быть и заполненными некоторыми звуками, в таком случае соответствующие буквенные символы ставятся внутри скобок: напр., (ы-ы).

В потоке речи в связи со спонтанностью и неподготовленностью устной речи нередко наблюдается употребление неправильных конструкций в грамматическом плане, повторов, вставных конструкций, самокоррекций, обрывов [14, с. 24]. Иначе говоря, неподготовленная устная речь самоорганизуется в процессе порождения, в отличие от речи письменной, где наблюдается только фиксированный результат [15, с. 27].

В процессе транскрибирования аудиофайлов вышеуказанные хезитационные явления учитываются и передаются специальными знаками.

Самокоррекция возникает в тех случаях, когда информант решает, что по той или иной причине определенный фрагмент порожденного им высказывания не соответствует тому, что он намеревался произнести. Информант немедленно реагирует на ошибку, например на неправильное слово в соответствующем контексте (лексико-семантические ошибки) или некорректную грамматическую форму (морфосинтаксические ошибки), и в последущем исправляет неверный элемент корректным с помощью замены: *Әтийемде (э-э) суғышҡа [алып кителгән] / алып киткәндәр март айында // ‘Отца [был отправлен] отправили на фронт в марте’*.

Особое место занимают в потоке речи обрывы (..), когда информант начинает произносить то или иное слово, по какой-то причине не закончив его, начинает вновь: *анан аҙаҡ (ы-ы) / уралдан бер # участок # алыб / ул уралдағы учасканы / (//) өй.. (//) / беҙҙең өйҙән ул бер + умбиш километр йер / урал аша үтәргә кирәк / урал аша / шул // ‘потом взяли на Урале покосный участок / до которого было около пятнадцати километров от нашего дома / и надо было перейти через Урал / вот //’*

Просодические характеристики речи, такие как ударение, интонация, тон, на данном этапе транскрибирования диалектного материала не учитывались.

Если в письменной речи структурно-смысловое членение высказывания осуществляется с помощью пунктуационных средств, то в устной речи звуковой поток членится на синтагмы (/) (смысловое целое, отделенное небольшой паузой) и фразы (//) (законченное целое, которое может состоять из группы синтагм, но может состоять и из одной синтагмы и которое нормально характеризуется конечным понижением тона).

Специальные знаки транскрибирования:

/ – знак сегментирования синтагмы;

// – знак сегментирования повествовательного высказывания (членение на фразы и синтагмы осуществляется с учетом интонационно-синтаксических характеристик отрезков звуковой цепи);

! ? – знак сегментирования вопросительных и восклицательных фраз;

(..) – знак длительной паузы в потоке речи информатора;

( ) – знак паузы хезитации, если она заполнена некоторыми звуками, соответствующие буквенные символы ставятся внутри скобок: напр., э (э-э);

(//) – обрыв высказывания перед последующим словом;

[ ] – самокоррекция (неправильное слово или грамматическая форма, которые в последующем исправляются информантом);

.. – обрыв слова перед последующим словом;

$ Н – не поддающееся расшифровке или неуверенно расшифрованное слово или словосочетание;

$ С – паралингвистический элемент речи – смех;

$ К – паралингвистический элемент речи – кашель;

$ В – паралингвистический элемент речи – вздох, стон, причмокивание и др.

... # – переключение языковых кодов (с башкирского на русский);

+ – для указания таких языковых явлений, как элизия и ассимиляция (+баралманы – бара алманы, +ҡайтыб*ара – ҡайтып бара);

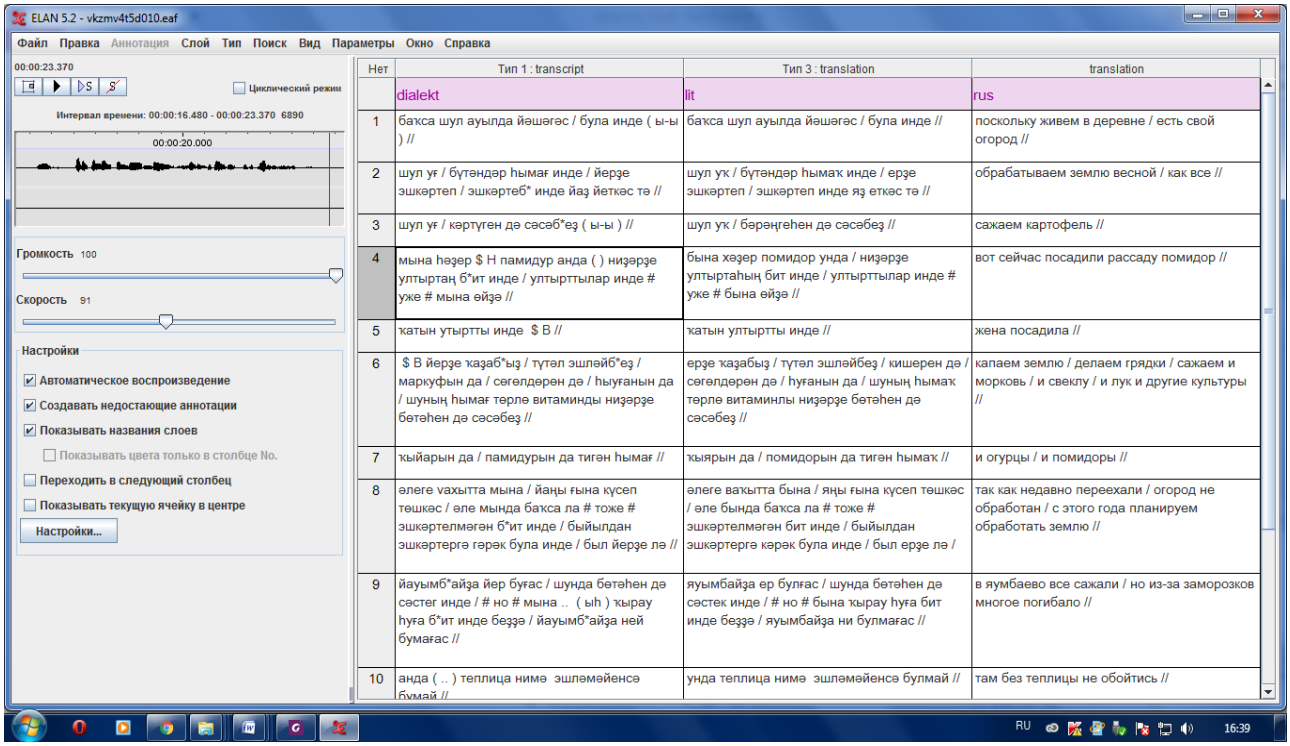

Рис. 2. Подготовительная работа диалектного материала в программе ELAN

В диалектном корпусе такая многоуровневая разметка, включающая и паралингвистические элементы, хезитационные явления, окажется бесценным источником при исследовании языка с точки зрения когнитивных, социологических и психологических подходов и повысит информативность корпуса и расширит круг задач, решаемых с помощью корпуса.

5. Заключение. На сегодняшний день обработаны аудиофайлы 30 информантов по Бурзянскому, Учалинскому и Дуванскому районам Республики Башкортостан. Предстоит сбор полевых материалов по населенным пунктам остальных районов, которые были определены как опорные для построения Диалектологического атласа башкирского языка.

В дальнейшем планируется разработка корпуса-менеджера для осуществления поисковых запросов по транслитерации и литературному эквиваленту.

Корпус восточного диалекта башкирского языка предоставит возможность свободного доступа лингвистов различной специализации к первичному диалектному материалу, выступит в качестве бесценного источника для изучения отдельного говора, установления территории распространения того или иного языкового явления, станет базой для изучения исторического развития и становления литературного языка, социолингвистического анализа, сравнительно-сопоставительных исследований языков.

Литература

1. Москвина Т.Н. Методы и подходы корпусной лингвистики в исследованиях семантики диалектной лексики // Современные проблемы науки и образования. 2014. № 6. URL: // http://science-education.ru/ru/article/view?id=15784 (дата обращения: 18.06.2019).

2. Крючкова О.Ю., Гольдин В.Е. Корпус русской диалектной речи: концепция и параметры оценки // Компьютерная лингвистика и интеллектуальные технологии: Мат-лы Международной конференции «Диалог – 2011». Бекасово, 2011. С. 359–367.

3. Максютова Н.Х. Восточный диалект башкирского языка (в сравнительно-историческом освещении). М., 1976.

4. Шакуров Р.З. Диалектная система башкирского языка // Ватандаш. 2012. № 8. С. 40–61.

5. Словарь башкирских говоров. Восточный диалект / Ред. Н.Х. Максютова. Т.1. Уфа, 1967 (на башк. яз.).

6. Диалектологический словарь башкирского языка. Уфа, 2002 (на башк. яз.).

7. Диалектологический атлас башкирского языка / Под ред. Ф.Г. Хисамитдиновой. Уфа, 2005.

8. Образцы башкирской разговорной речи / Ред. Н.Х. Максютова. Уфа, 1988.

9. Кибрик А.А., Подлесская В.И. К созданию корпусов устной русской речи: принципы транскрибирования // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2003. № 10. С. 5–13.

10. 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 // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 7 (14). М., 2008. С. 488–495.

11. Куканова В.В., Бембеев Е.В., Убушаев Н.Н., Манджиева Б.Б. Устные тексты на калмыцком языке: запись и расшифровка // Вестник Калмыцкого университета. 2013. №3 (19). С. 56–64.

12. Блинова О.В. Будешь задумываться / что говорить / а что не говорить: парадокс наблюдателя и речевой контроль в ОРД. URL: // https://www.slm.uni-hamburg.de/slavistik/forschung/veranstaltungen/symposium-sprachvariation/downloads-symposium-april-2016/blinova.pdf (дата обращения: 18.06.2019)

13. Белицкая А.А. О роли хезитационных пауз в спонтанной речи // Филология и литературоведение. 2014. № 2. URL: // http://philology.snauka.ru/2014/02/697 (дата обращения: 18.06.2019).

14. Бускунбаева Л.А. Закономерности речевой экономии и их отражение в башкирском языке. Уфа, 2008.

15. Звуковой корпус как материал для анализа русской речи. Ч. 1. Чтение. Пересказ. Описание / Под ред. Н.В. Богдановой-Бегларян. СПб., 2013.

References

1. Moskvina T.N. Methods and approaches of corpus linguistics in studying the semantics of dialectal vocabulary. Sovremennye problemy nauki i obrazovaniya, 2014, no. 6. Available at: http://science-education.ru/ru/article/view?id=15784 (accessed June 18, 2019).

2. Kryuchkova O.Yu., Goldin V.E. Corpus of Russian dialect speech: The concept and parameters of evaluation. Kompyuternaya lingvistika i intellektualnye tekhnologii. Proceedings of the International Conference «Dialogue-2011». Bekasovo, 2011, pp. 359–367.

3. Maksyutova N.Kh. The Eastern dialect of the Bashkir language in comparative historical consideration. Moscow, Nauka, 1976. 292 p.

4. Shakurov R.Z. The dialect system of the Bashkir language. Vatandash, 2012, no. 8, pp. 40–61.

5. Dictionary of Bashkir dialects. The Eastern dialect. Maksyutova N.Kh (ed.). Vol. 1. Ufa, 1967. 300 p.

6. Dialectological dictionary of the Bashkir language. Ufa, 2002. 432 p.

7. Dialectological atlas of the Bashkir language. F.G. Hisamitdinova (ed.). Ufa, 2005. 234 p.

8. Samples of Bashkir colloquial speech. Maksyutova N.Kh. (ed.). Ufa, 1988. 224 p.

9. Kibrik A.A., Podlesskaya V.I. On the creation of oral Russian speech corpora: Principles of transcription. Nauchno-tekhnicheskaya informatsiya. Ser. 2. Informatsionnye protsessy i sistemy, 2003, no. 10, pp. 5–13.

10. Stepanova S.B., Asinovsky A.S., Bogdanova N.V., Rusakova M.V., Sherstinova T.Yu. Sound corpus of the Russian language of "One Speech Day" everyday communication: Concept and state of formation. Kompyuternaya lingvistika i intellektualnye tekhnologii. Proceedings of the annual international conference “Dialogue”. Issue 7 (14). Moscow, 2008, 488–495 pp.

11. Kukanova V.V., Bembeev E.V., Ubushaev N.N., Mandzhiev B.B. Oral texts in the Kalmyk language: Records and decoding, Vestnik Kalmytskogo universiteta, 2013, no. 3 (19), 56–64 pp.

12. Blinova O.V. You will think / what to say / what not to say: The observer’s paradox and voice control in "One Speech Day". Available at: www.slm.uni-hamburg.de/slavistik/forschung/veranstaltungen/symposium-sprachvariation/downloads-symposium-april-2016/blinova.pdf (accessed June 18, 2019).

13. Belitsaya A.A. On the role of hesitation pauses in spontaneous speech. Filologiya i literaturovedenie, 2014, no. 2. Available at: http://philology.snauka.ru/2014/02/697 (accessed June 18, 2019).

14. Buskunbaeva L.A. Regularities of speech economy and their reflection in the Bashkir language. Ufa, 2008. 139 p.

15. The sound corpus as a material for the analysis of Russian speech. Ch. 1. Reading. Retelling. Description. N.V. Bogdanova-Beglaryan (ed.). St. Petersburg, 2013. 532 p.

CREATING THE AUDIO CORPUS OF THE EASTERN DIALECT OF THE BASHKIR LANGUAGE: PROBLEMS AND PROSPECTS

© L.A. Buskunbaeva, Z.A. Sirazitdinov

The Badge of Honour Order Institute of History, Language and Literature,
Ufa Federal Research Centre, Russian Academy of Sciences
71, prospekt Oktyabrya, 450054, Ufa, Russian Federation

This article discusses the principles of creating a corpus of audio materials concerning the Eastern dialect of the Bashkir language. The task is to representatively collect field materials taking into account age, gender, level of education, language of communication, national identity of the informers and diversity of topics for conversation.

The corpus assumes the presentation of audio recording, decryption of sound files in the form of transcribed texts, their literary version and Russian translation of the dialect text.

The authors propose a transcription in "semi-orthographic recording" most widely used in speech corpora. The developed system of graphic transmission of dialect features is based on special characters and letter combinations. The division of the informer's thematic monologue into phrases and syntagms, transcription and synchronization of the transcript with the audio file is performed using the ELAN annotation program. Prosodic features, such as stress, intonation and tone, are not taken into account at this stage of transcribing the dialect material.

Key words: Bashkir language, Turkic languages, linguistic corpus, corpus linguistics, dialectology, expeditionary material.