МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ
УФИМСКИЙ ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР
РОССИЙСКОЙ АКАДЕМИИ НАУК
(УФИЦ РАН)
ОРДЕНА ЗНАК ПОЧЕТА ИНСТИТУТ ИСТОРИИ, ЯЗЫКА И ЛИТЕРАТУРЫ-
ОБОСОБЛЕННОЕ СТРУКТУРНОЕ ПОДРАЗДЕЛЕНИЕ УФИМСКОГО ФЕДЕРАЛЬНОГО
ИССЛЕДОВАТЕЛЬСКОГО ЦЕНТРА РОССИЙСКОЙ АКАДЕМИИ НАУК
(ИИЯЛ УФИЦ РАН)

УДК 512.141:004.655
Рег. № НИОКТР АААА-А17-117040710023-4

СОГЛАСОВАНО
Директор ИИЯЛ УФИЦ РАН
д.г.н., проф.
_____ А.В. Псянчин
«16» апреля 2020 г.

УТВЕРЖДАЮ
Врио председателя УФИЦ РАН
д.х.н., проф.
_____ А.Г. Мустафин
«16» апреля 2020 г.

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

СОЗДАНИЕ КОРПУСА ДИАЛЕКТНЫХ ТЕКСТОВ БАШКИРСКОГО ЯЗЫКА

(промежуточный, этап 3)

Руководитель НИР,
к.ф.н., в.н.с. _____ З.А. Сиразитдинов

Уфа 2020

Список исполнителей

Руководитель темы:

Кандидат филологических наук, ведущий научный сотрудник

16.0.2020 З.А.Сиразитдинов

Исполнители темы:

16.01.2020 1) к.ф.н., в.н.с. З.А.Сиразитдинов,

16.01.2020 2) к.ф.н., с.н.с. Л.А.Бускунбаева,

16.01.2020 3) м.н.с. А.Ш.Ишмухаметова,

16.01.2020 4) м.н.с. Г.Г.Шамсутдинова

Отчет о НИР в объеме 31страницы по теме «СОЗДАНИЕ КОРПУСА ДИАЛЕКТНЫХ ТЕКСТОВ БАШКИРСКОГО ЯЗЫКА»

Промежуточный отчет. Отражает основное содержание и результаты работ за 2019 г.

РЕФЕРАТ

КЛЮЧЕВЫЕ СЛОВА: филология, тюркология, башкирский язык, корпусная лингвистика, расшифровка речи, фонетическая транскрипция, диалектология, башкирские говоры, лингвистические базы данных.

ОБЪЕКТ ИССЛЕДОВАНИЯ. Объектом данного научного исследования являются говоры восточного диалекта башкирского языка.

ЦЕЛЬ РАБОТЫ. Целью научной работы является создание речевого корпуса по говорам восточного диалекта башкирского языка.

МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ. При работе над проектом используется методология сбора и компьютерной обработки полевых речевых материалов информаторов.

РЕЗУЛЬТАТЫ РАБОТЫ И ИХ НОВИЗНА. Собран и обработан полевой речевой материал по современному состоянию говоров восточного диалекта башкирского языка. Полевые аудиоматериалы паспортизованы и транскрибированы в соответствии с разработанными авторами принципами. Разработана среда представления транскрибированных текстов в информационной системе Машинный фонд башкирского языка (МФБЯ).

Транскрибированные материалы экспедиционных аудиозаписей по говорам восточного диалекта введены в базу данных Машинного фонда башкирского языка.

ОБЛАСТЬ ПРИМЕНЕНИЯ. Результаты научно-исследовательской работы найдут применение в диахронных и синхронных исследованиях башкирских

языковедов, в историко-сравнительных научных изысканиях тюркологов и сопоставительных исследованиях алтаистов. Обработанные полевые материалы и база данных транскрипция фонетической речи восточного диалекта могут использоваться в учебном процессе – при чтении спецкурсов, лекций по фонетике и диалектологии башкирского языка, при написании учебных пособий для ВУЗов. Результаты могут быть полезными при создании и реализации проектов, направленных на сохранение языка и культуры носителей говоров национального языка. Созданный материал может быть использован в решении смежных с филологическим вопросов по культуре, этнографии, фольклору, искусствоведению, синтезу и анализу башкирской речи.

1. ВВЕДЕНИЕ … … … … … … … … … … … … . … … … … … . . . . . … … . … . 6
2. ОСНОВНАЯ ЧАСТЬ… … … … … … … … … … … … … . … … … . . . . . … … . 8
2.1. СБОР ПОЛЕВОГО МАТЕРИАЛА… … … … … … … … … … . . . … . … . 8
2.2. ПРИНЦИПЫ ОБРАБОТКИ И ТРАНСКРИБИРОВАНИЯ ПОЛЕВОГО МАТЕРИАЛА……………………………………..… … … . . . . … … . 1 1
2.3. СРЕДСТВА ПРЕДСТАВЛЕНИ КОРПУСНЫХ ДАННЫХА ДАННЫХ.. 17
2.4. ПРОВЕДЕННАЯ РАБОТА … … … … … … … … . … … … . … . . . … … … . 20
3. ЗАКЛЮЧЕНИЕ… … … … … … … . … … … … … … … … . . . . … … … 22
4. ИСТОЧНИКИ И ЛИТЕРАТУРА … … . … … … … … … … . . . . … … … 23

ВВЕДЕНИЕ

Приоритетным направлением мировой и отечественной лингвистики является корпусная лингвистика, объектом исследования которой являются речевые материалы, реализованные как в виде письменных текстов, так и устных (фонетических) массивов данных, снабженных специальной системой разметок. По объему и в количественном соотношении письменные корпусы намного превосходят устные. Например объемы корпусов немецкого, английского китайского язуу измеряются миллиардами слов [1-7], но в то же время темп составления речевых корпусов с каждым годом растет, что определяется все более возрастающим интересом исследователей к разговорному тексту, анализу особенностей функционирования языковых единиц в речевом потоке. Среди аудикорпусов выделяются корпусы немецкого, английского, китайского, турецкого, японского языков [8-12].

В отечественной филологии описаны речевые корпусы ISABASE (ИСА РАН и филологический ф-т МГУ) [13], RuSpeech (ИСА РАН и корпорация Intel) [14, 15], TeCoRus (ВЦ РАН и МГЛУ) [16]. Однако эти корпусы не представлены в сети Интернет в открытом доступе. Однако в сети в открытом доступе имеются фонетические корпусы “Один день русского языка” [17], “Рассказы о сновидениях”, «Рассказы сибиряков о жизни», «Весёлые истории из жизни» [18], представляющие не меньший научный интерес.

Инновационные 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 функционирует в составе национальных корпусов. Так Национальный корпус русского языка включает в себя подкорпус диалектных текстов, который функционирует с существует с 2005 года [19]. Данный подкорпус представляет аудиорасшифровки полевых материалов в орфографизированном виде, дает информацию об особенностях диалектной грамматики, приводит «паспорт» записи (место, время, данные об информанте и лице, предоставившем текст), имеются указания на жанрово-тематическую отнесенность текстов, на «диалектность» значений отдельных слов. Диалектологический подкорпус НКРЯ в переспективе предполагает включение любых диалектных текстов на русском языке, записанных как на территории исконного проживания русского населения (Европейская часть России), так и на территориях раннего заселения (Русский Север), позднего заселения (Сибирь, Дальний Восток, Дон, Нижнее Поволжье) и миграций (говоры старообрядцев протестантов Латгалии, Азербайджана, Румынии, Австралии, Канады, Америки); аудиорасшифровки, тексты из малотиражных сборников и изданий, войдут также видеорасшифровки и тексты из хрестоматий [20-21].

Силами преподавателей и студентов Томского университета создается “Томский диалектологический корпус”, представляющий расшифровки русской речи сибиряков [22].

Графическая запись звуковой речи позволяет получить значительную информацию по диалектной речи, однако, как отмечают исследователи “никакая, даже самая тщательная, транскрипция не способна заменить звучащий текст, передать все его богатство” [18]. Поэтому сейчас возрастает интерес к сопровождению графическай нотации аудиоматериалами. Такой подход наблюдается в разработках Саратовского диалектологического корпуса [23] Мультимедийного корпуса диалектных текстов Устьянского района Архангельской области [24], мультимедийного корпуса ижемского говора коми языка [25].

Среди отечественных языковедов активно ведутся работы по созданию речевых корпусов калмыцкого [26], бурятского [27], тувинского [28] и других языков народов Российской Федерации.

Следует особо отдельно отметить уникальный проект, разрабатываемый учеными Института языкознания РАН 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 [29].

Разработка корпусов диалектных текстов сопряжена с целым рядом сложностей, куда 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 [30]. 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» [20:363].

ОСНОВНАЯ ЧАСТЬ

2.1. СБОР ПОЛЕВОГО МАТЕРИАЛА

Создание репрезентативного диалектного корпуса требует максимального охвата всех говоров и содержательной металигвистической 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. В 2017 г. был собран материал по кызыльскому говору восточного диалекта.

2) По каждому населенному пункту предполагается осуществить запись по гендерному признаку, и 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 (что любят готовить на каждый день, что на праздниках);

– какие домашние животные держат и как содержат;

– дети и близкие родственники (система родства);

– приусадебное хозяйство (огород, сад);

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

– топонимия в окрестностях поселения;

– история села, школы, рода;

– работа, школа, друзья;

– времена года, погода;

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

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

– игры детей;

– друзья детей (какие у детей друзья, 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 производится в программе Sound Forge.

7) Паспортизация файлов и экстралингвистическая разметка осуществляется в базе данных Access, частично включается 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 указывается 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.

В ходе работы выяснилась необходимость представления паспортизации аудиофайлов в виде базы данных Access с указанием длительности записи и некоторых дополнительных сведений об информаторе. На рис.1 представлен скрин-шот структуры базы.

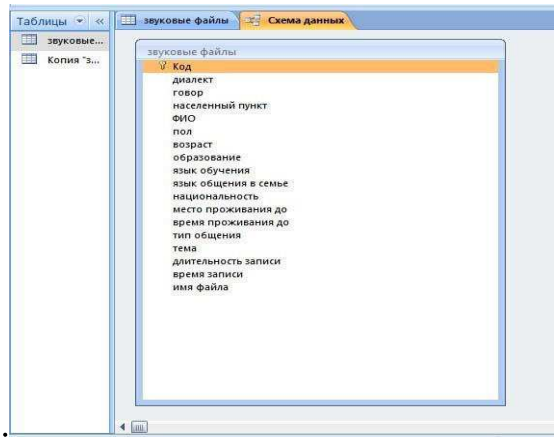

Рис. 1. Схема данных паспортизации файлов и экстралингвистической разметки в базе Access.

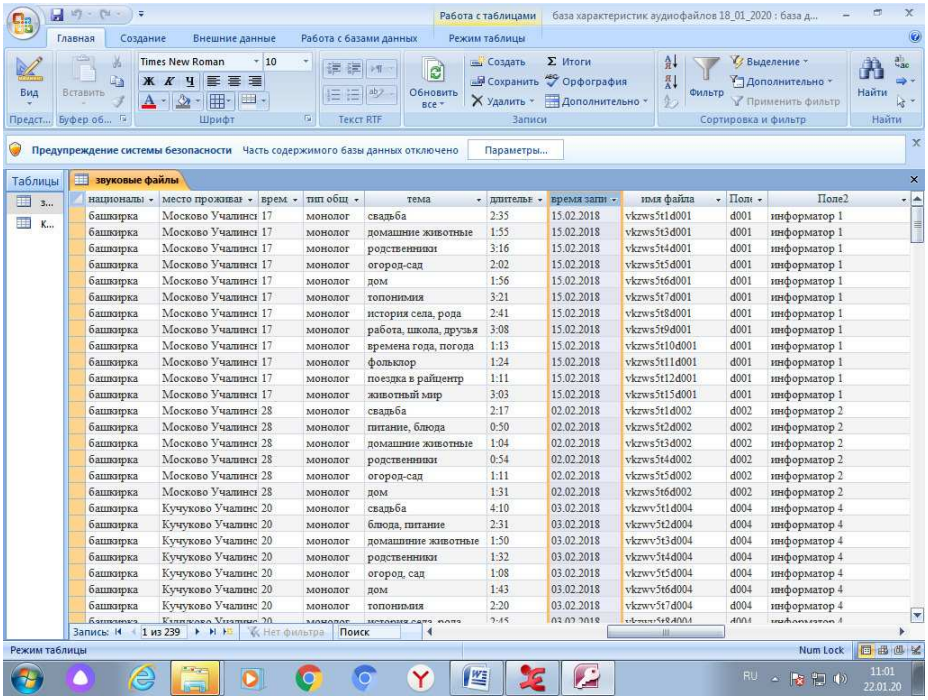

Рис.2. Скриншот из базы Access паспортных данных полевых материалов.

2.2. ПРИНЦИПЫ ТРАНСКРИБИРОВАНИЯ И ЛИНГВИСТИЧЕСКОЙ РАЗМЕТКИ

В корпусе представляются аудиозаписии, расшифровки звуковых файлов в виде транскрибированных текстов, их литературный вариант и русский перевод диалектного текста.

Членение аудиофайла на коммуникативные эпизоды (фразы), транскрибирование и синхронизация транскрипта с аудиофайлом осуществляется 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 [21: 4]. Эта транскрипция близка к фонематической, максимально приближена к орфографии современного башкирского языка. Разработчиками была разработана единая система графической передачи диалектных особенностей с использованием специальных знаков и буквосочетаний для обозначения тех звуков, которых 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 (ҡ ара – ҡ а*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, h выделены следующие согласные:

v – билабиальная шумная согласная в (ваҡ ыт – vаҡ ыт, вәғәҙ ә – vәғәҙ ә),

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

ҡ * – гортанный вариант согласной [ҡ] (ҡ алын – 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.

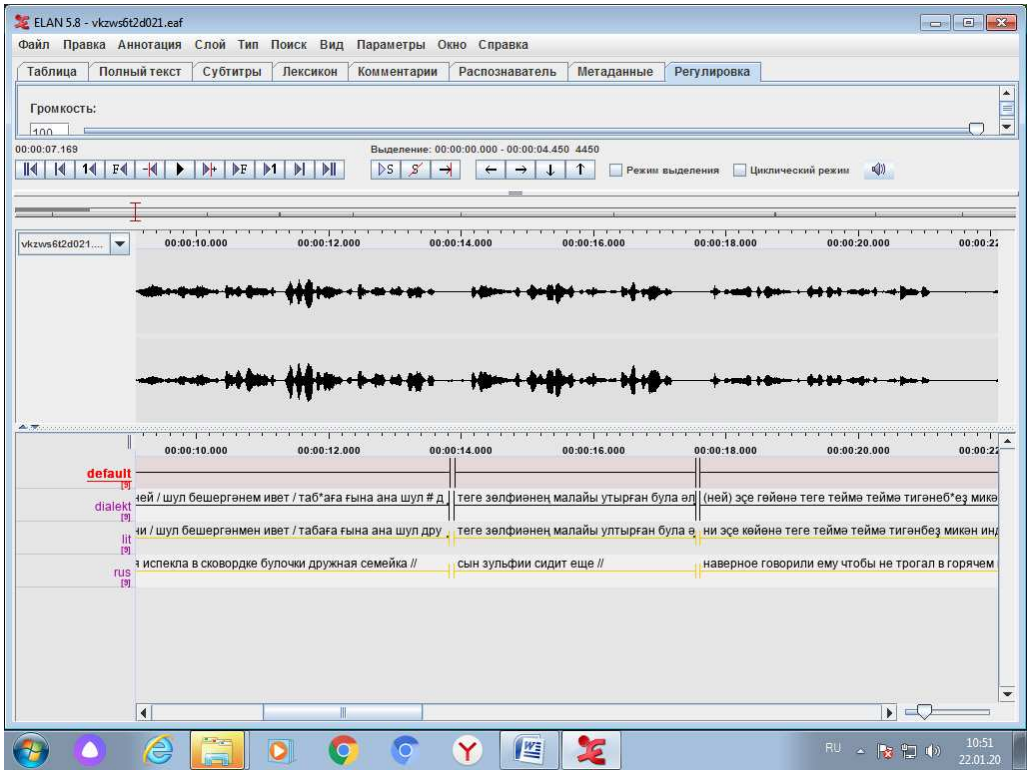

Рис.3. Обработка аудиоматериала в программе ELAN (транскрибирование диалектного материала, создание литературного варианта, перевод на русский язык)

Надо отметить, что вышеуказанная транскрипция касается 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) нейҙ әрҙә беҙ ҙең (э-э) буш ʋакытында / теге # пасиф # ʋакыты бөттө нисә / китереб беҙ ҙең капка төб*өнә өйөб куйа 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 др., хезитационные явления, которые сопровождают живую речь.

При разработке аннотирования на уровне высказываний в орфографической записи был применен опыт создателей диалектологических и мультимедийных корпусов русского и калмыцкого языков [22-24].

Особое внимание уделяется и непреднамеренным остановкам информационного потока в процессе коммуникации, которые обусловлены целым рядом факторов как индивидуальных, психологических, так и физиологических. Поскольку как заметил американский паузолог О'Коннор, «паузы могут сказать о человеке не меньше, чем слова и в разговоре тратится 40–50% времени на них» [25]. 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» [25].

В потоке речи информатора 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 [26: 24]. Иначе говоря, неподготовленная устная речь самоорганизуется в процессе порождения, в отличие от речи письменной, где наблюдается только фиксированный результат [27: 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 (морфо-синтаксические ошибки), и в последущем исправляет неверный элемент корректным с помощью замены: *Әтийемде (э-э) суғышқа [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 ее, начинает вновь: *анан аҙ ақ (ы-ы) / уралдан бер # участок # алыб / ул уралдағы участканы / (//) өй.. (//) / беҙ ҙ ең өйҙ ән 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, как ударение, интонация, тон, на данном этапе транскрибирования диалектного материала не учитывались.

Языковые явления, такие как элизия и ассимиляция, то есть изменения, происходящие в потоке речи на стыке слов, тоже выделяются специальным знаком (+).

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» [28: 87]).

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 языковых явлений, как элизия и ассимиляция (+баралманы — бара алманы, +к айтыб*ара – к айтып бара);

В диалектном корпусе такая многоуровневая разметка, включающая и паралингвистические элементы (кашель, смех, вздох), 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.

Речевой материал разбивается на фразовые промежутки, которые в дальнейшем транскрибируются в слое транскрипция и связываются со слоями

представления диалектной фразы на современном башкирском литературном языке и переводом на русский язык. В дальнейшем это даст возможность включить материал в самостоятельный диалектный корпус башкирского языка с реализацией поиска по словарной и морфологической базам.

2.3. СРЕДСТВА РАБОТЫ С КОРПУСОМ АУДИО- И ТЕКСТОВЫХ ДАННЫХ

Для паспортизированного и транскрибированного текстового материала аудизаписей разработана возможность представления в Машинном фонде башкирского языка.

Машинный фонд башкирского языка — это специально разработанная для лингвистов, преподавателей, студентов и учащихся старших классов система поиска филологической информации. Данная информационная система включает в себя специализированные базы данных, интерфейсы для выполнения запросов к базам данных и программные обеспечения для обработки этих запросов. На рис. 4 представлена структурная схема информационной системы.

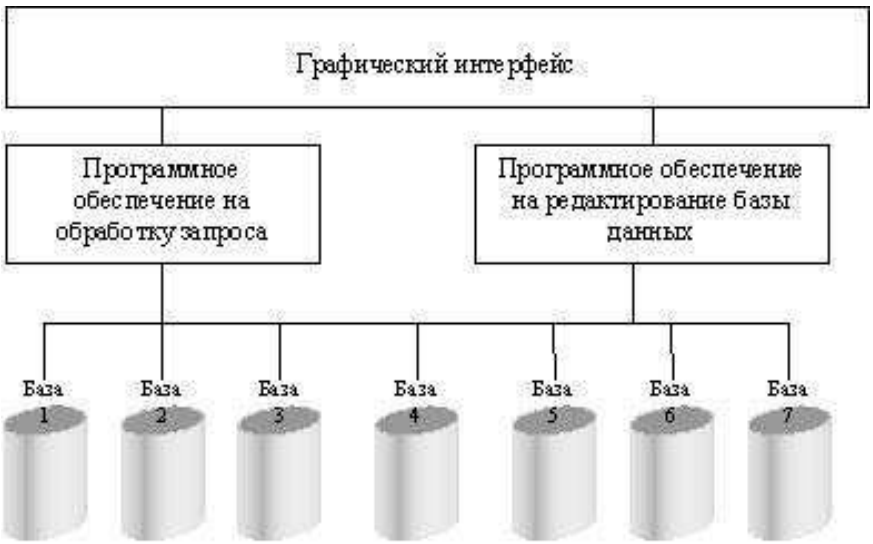

Рис. 4. Структура Машинного фонда башкирского языка.

Фонд разработан на базе СУБД Oracle и функционирует на специально выделенном сервере по адресу: www.mfbl2.ru.

На сегодняшний день эта система имеет 7 крупных баз данных, которые образуют подфонды единого Машинного фонда:

- подфонд генеральной картотеки;
- лексикографический подфонд;
- грамматический подфонд;
- подфонд каталога рукописных книг;
- подфонд каталога старопечатных книг;
- экспериментально-фонетический подфонд;
- диалектологический подфонд.

В диалектологическом подфонде разработан раздел для представления данных (структура базы данных транскрипций аудиофайлов), средства и интерфейсы для ввода транскрибированного материала, многофункционального поиска и вывода текстового материала для конечного пользователя.

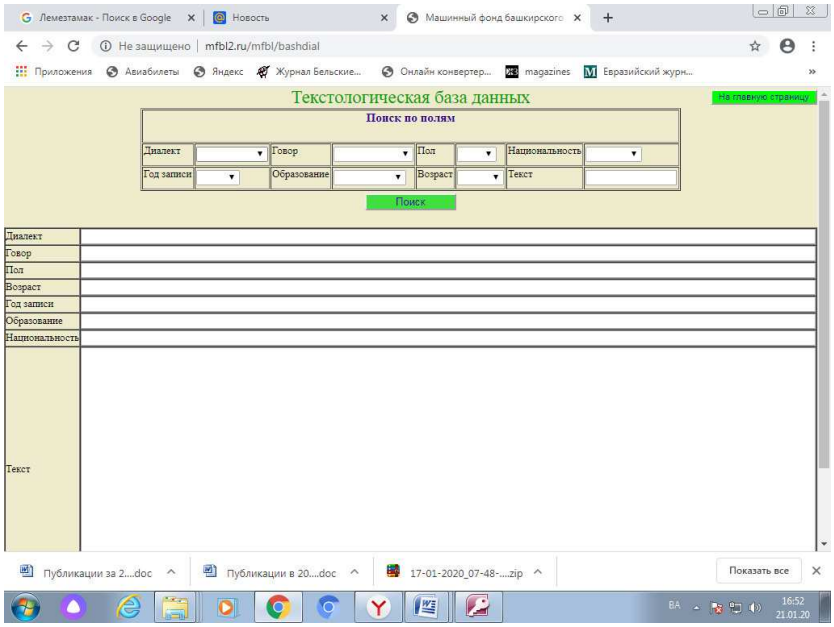

Рис.5. Интерфейс для ввода транскрипций в базу Машинного фонда.

На рис. 5. представлен интерфейс для ввода паспортных данных и транскрипций в диалектологическом подфонде Машинного фонда башкирского языка: администраторы вводят в поля паспортные данные текста, в поле “текст”

— транскрипцию текста из базы данных программы ELAN(скопировав его в буфер памяти и затем вставив в окне “текст” на рис. 3).

Средства поиска позволяют искать и выдавать тексты по следующим паспортным параметрам: диалекту, говору, году записи, образованию, полу, национальности и возрасту информатора.

На рис. 6. представлена выдача текста по следующим параметрам: восточный диалект, кызыльский говор, год записи – 2018, образование информатора – высшее, возраст 57 лет, национальная идентификация – башкир.

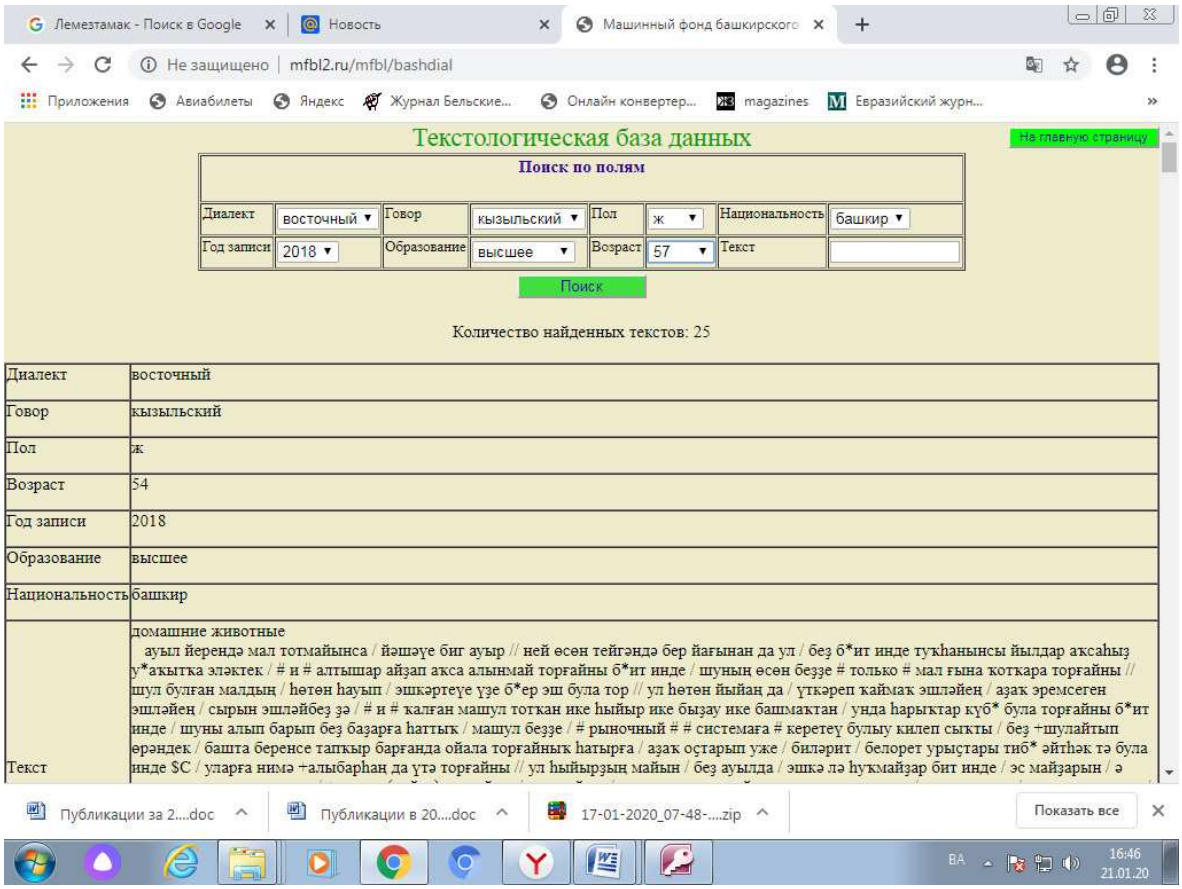

Рис. 6. Пример выдачи транскрипции аудиофайлов по выбранным параметрам поиска.

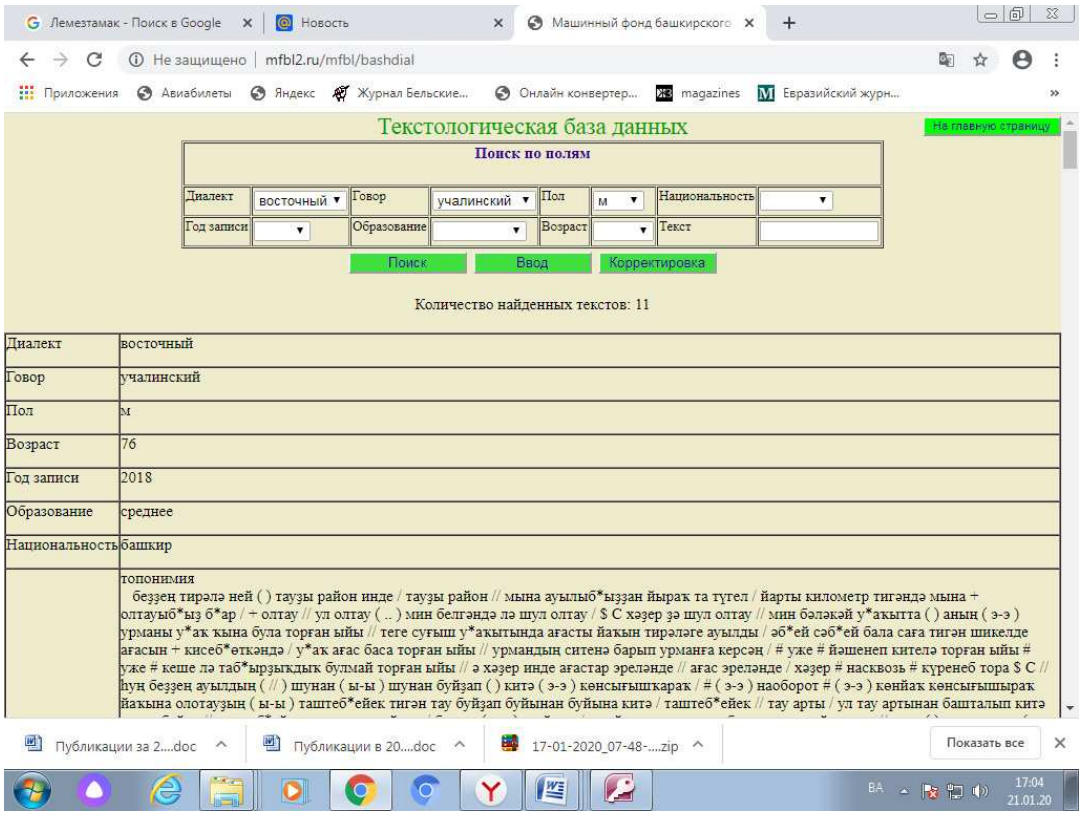

Рис.7. Пример выдачи транскрипций аудиофайлов информаторов мужчин учалинского говора.

2.4.ПРОВЕДЕННАЯ РАБОТА

1) Вследствии отсутствия финансирования экспедиционных поездок, запланированные записи информаторов по сальютскому, миасскому и аргаяшского говорам не были выполнены. Вместо этого были осуществлены дополнительные записи по кызыльскому, учалинскому и айскому говорам за счет личных поездок исполнителей темы в места проживания близких родственников. Всего собран за отчетный год материал от 20 информаторов, представляющих 210 записей по заданным темам объемом в 1,9 Гбайт.

2) Проведена первичная обработка полевого материала на удаление шумов и длительных пауз в программе SoundForige.

3) Осуществлена сегментация цифрового речевого материала. Данная процедура проведена в программе ELAN и включает сегментирование речевого потока на фразы.

4) Проведено транскрибирование цифрового материала. Транскрибирование осуществлено в программе ELAN по разработанной нами методике транскрибирования полевого материала.

5) Осуществлено аннотирование цифрового материала. Аннотирование проведено в программе ELAN. В данную процедуру включено аннотирование таких фонетических явлений как элизия, протеза, выделены синтагматические единицы.

6) Проведена паспортизация оцифрованного полевого материала.

7) Осуществлен ввод 240 транскрибированных за предыдыущие годы текстов кызыльского, учалинского и айского говоров из формата ELAN в текстовую базу данных диалектологического подфонда Машинного фонда башкирского языка (mfbl2.ru).

9) Начат ввод материалов в проект Института языкознания РАН [http://lingvodoc.tsu.ru/].

10) Опубликовано 30 научных публикаций [40-70].

ЗАКЛЮЧЕНИЕ

За время работы в 2019 г. доработаны принципы подготовки и обработки полевого материала. Обработан речевой материал экспедиций по трем говорам: кызыльский, учалинский и айлинский. Транскрибированный материал представлен в разделе текстологическая база диалектологической базы данных Машинного фонда башкирского языка. Идет представление фонетических аудиофайлов в информационной базе “Лингводок”.

В рамках научного проекта коллективом опубликовано 30 научных публикаций [40-70]. Из них 6 в рецензируемых журналах, 1 коллективная монография в соавторстве с зарубжными учеными.

4. ИСТОЧНИКИ ЛИТЕРАТУРА

1. Das Deutsche Referenzkorpus (DeReKo) URL:http://www.ids-mannheim.de/kl/projekte/korpora (дата обращения: 25.12.2019).

2. British National Corpus (BNC). URL:http://www.natcorp.ox.ac.uk (дата обращения: 25.12.2019).

3. The corpus of contemporary american english (COCA). URL:http://corpus.byu.edu/coca (дата обращения: 25.12.2019).

4. Корпус китайского языка. (LIVAC Synchronous Corpus). URL:http://www.rcl.cityu.edu.hk/livac (дата обращения: 25.12.2019).

5. Magyar Nemzeti Szövegtár (корпус венгерского языка). URL:http://corpus.nytud.hu/mnsz (дата обращения: 25.12.2019).

6. Corpus del español (корпус испанского языка). URL:http://www.corpusdelespanol.org (дата обращения: 25.12.2019).

7. Национальный корпус русского языка (НКРЯ) URL:http://www.ruscorpora (дата обращения: 25.12.2019).

8 LDC Top Ten Corpora (мультимедийные корпусы английского языка). URL: http://www.ldc.upenn.edu/Catalog/topten.jsp (дата обращения: 17.01.2020).

9. Фонетический немецкого разговорного языка URL:http://dsav-wiss.ids-mannheim.de/korpora/pf/pf _doku.htm (дата обращения: 22.01.2020).

10. Chinese Broadcast Conversation Speech (мультимедийный корпус китайского языка). URL: http://www.ldc.upenn.edu/Catalog/ CatalogEntry.jsp? catalogId= LDC2013S04 (дата обращения: 17.01.2020).

11. The *Corpus* of Spontaneous Japanese (мультимедийный корпус японского языка). URL: http://www.ninjal.ac. jp/products-k/katsudo/seika/corpus/public /index.html (дата обращения 25.01.2020).

12. *The Spoken Turkish Corpus (*мультимедийный корпус разговорного турецкого языка). URL:http://stc.org.tr (дата обращения: 22.01.2020).

13. Богданов Д.С., Кривнова О.Ф., Подрабинович А.Я., Фарсобина В.В. База речевых фрагментов русского языка ISABASE // Сб. «Интеллектуальные технологии ввода и обработки информации». М., Эдиториал УРСС, 1998.

14. Богданов Д.С., Брухтий А.В., Кривнова О.Ф., Подрабинович А.Я., Строкин Г.С. Технология формирования речевых баз данных // Сб. «Организационное управление и искусственный интеллект». М., Эдиториал УРСС, 2003.

15. Кривнова О.Ф. Области применения речевых корпусов и опыт их разработки/ Труды XVIII сессии Российского акустического общества РАО. Таганрог, 2006. С. 81–84.

16. Речевой корпус данных TeCoRus/ Электронный ресурс. URL: http://www.ccas.ru/depart/chuchu/doc_ru/projects/Tecorus_info_ru.pdf/ (дата обращения: 22.01.2020).

17. Фонетический корпус “Один день русского языка” URL:http://www.model.org.spbu.ru (дата обращения: 17.112.2019, доступ ограничен).

18. Рассказы о сновидениях и другие корпуса звучащей речи. Электронный ресурс. URL:http://www. spokencorpora.ru/ (дата обращения: 17.12.2019).

19. Корпус Диалектных Текстов в Национальном корпусе русского языка. Электронный ресурс. URL:http://www.ruscorpora.ru/search-dialect.html. (дата обращения: 17.12.2019).

20. *Качинская И.Б.* 2009. Корпус Диалектных Текстов в Национальном корпусе русского языка: состояние и перспективы // Лексический атлас русских народных говоров (Материалы и исследования). 2009. СПб. С. 57-68.

21. *Сичинава Д.В., Качинская И.Б.* 2014. Корпус диалектных тестов в национальном корпусе русского языка: сегодняшнее состояние и перспективы

//Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4-8 июня 2014 г.). Вып. 13 (20). – М.: Изд-во РГГУ, 2014. С. 593-600.

22. Томский диалектный корпус. Электронный ресурс http://losl.tsu.ru/corpus (дата обращения: 09.01.2020).

23. Саратовский мультимедийный диалектный корпус. Электронный ресурс URL: ruscorpora.ru/sbornik2005/13letuchy.pdf. (дата обращения: 09.01.2020).

24. Мультимедийный корпус диалектных текстов Устьянского района Архангельской области URL: http:http://www.slavist.de/Pushkino/ (дата обращения: 22.01.2020).

25. Корпус ижемского диалекта коми языка URL:http://videocorpora.ru/ru/allmedia (дата обращения: 25.12.2019).

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

27. Абаева Ю. Д. Фонетическая база данных бурятских диалектов: принципы составления и разметки//Филологические науки. Вопросы теории и практики, № 1 (43) 2015, часть 1. С.13-16.

28. Электронный корпус текстов тувинского языка. Диалектный корпус тувинского языка. Электронный ресурс. URL:http://www.tuvancorpus.ru/?q =node/80 (дата обращения: 25.12.2019).

29. Информационная система Лингводок. URL: http:http://lingvodoc. ispras.ru. (дата обращения: 25.12.2019).

30. Москвина 2014 – *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: 22.01.2020).

31. Крючкова О.Ю., Гольдин В.Е. Текстовый диалектологический корпус как модель традиционной сельской коммуникации [Электронный ресурс]. URL: http://www.dialog-21.ru/digests/dialog2008/materials/html/41.htm (дата обращения: 20.01.2020).

32. 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.

33. 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» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14). – М.: 2008. С. 488–495.

34. Блинова О. В. Будешь задумываться / что говорить / а что не говорить: парадокс наблюдателя и речевой контроль в ОРД [Электронный ресурс]. URL: https://www.slm.uni-hamburg.de/slavistik/forschung/veranstaltungen/symposium-sprachvariation/downloads-symposium-april-2016/blinova.pdf. (дата обращения: 20.01.2020).

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

36. Белицкая А.А. О роли хезитационных пауз в спонтанной речи // Филология и литературоведение. 2014. № 2 [Электронный ресурс]. URL: http://philology.snauka.ru/2014/02/697 (дата обращения: 20.01.2020).

37. Бускунбаева Л. А. Закономерности речевой экономии и их отражение в башкирском языке. Уфа: Гилем, 2008. 140 с.

38. Звуковой корпус как материал для анализа русской речи. Ч. 1. Чтение. Пересказ. Описание /отв. ред. Н. В. Богданова-Бегларян. СПб.: Филол. ф-т СПбГУ, 2013. 532 с.

39. Щерба Л. В. Фонетика французского языка. Очерк французского произношения в сравнении с русским: пособие для студентов факультетов иностранных языков. М.: Изд-во лит. на иностр. языках, 1953 (1937). 257 с.

40. Сиразитдинов З.А., Ильясова Г.М. А.Кайдаров – новатор-тюрколог казахской лингвистики// Проблемы Востоковедения, №4, 2019, с.75-96.1.

41. Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш. Об обработке звуковых материалов для диалектологического аудиокорпуса башкирского языка//Turkologia (Казахстан, Туркестан), №4(96), 2019, С. 35-45.

42. Сиразитдинов З.А., Бускунбаева Л.А., Садыков Т.С. О принципах создания фонетических корпусов тюркских языков: на примере материалов говоров башкирского языка// Alatoo Academic Studies (Кыргызстан, Бишкек). 2019. № 3, С. 95-105.

43. Сиразитдинов З.А. Парные сочетания глагольных словоизменительных аффиксов башкирского языка// Alatoo Academic Studies (Кыргызстан, Бишкек). 2019. № 3, с. 106-114.

44. Сиразитдинов З.А., Боҫ ҡонбаева Л.А., Ишмөхәмәтова А.Ш., Шәмсетдинова Г.Ғ. Башҡорт теленең диалект аудиокорпусын төҙөү: проблемалар һәм перспективалар // Ватандаш. №12, 93–102 с.

45. Хисамитдинова Ф.Г., Бускунбаева Л.А., Ишкильдина Л.К., Муратова Р.Т., Ягафарова Г.Н. О тенденциях и перспективах изучения синтаксиса башкирского языка // Acta Linguistica Petropolitana. Санкт-Петербург, 2019, № 1. С. 363-404.

46. Жубанов А.Х., Садыков Т.Ш., Сиразитдинов З.А. Синтез тюркской словоформы. Алматы, 2019, 248 с.

47. Сиразитдинов З.А. Информационные лингвистические ресурсы башкирского языка и их роль в сохранении и расширении функций языка//Сохранение и развитие языков народов мира и развитие языкового разнообразия в киберпространстве: контекст, политики, практика: Материалы международной конференции. Якутск, 30 июня – 5 июля, 2019. С.

48. Сиразитдинов З.А. Об использовании лингвистического корпуса башкирского языка в переводческой деятельности //Проблемы полиязычного образования в современном Казахстане: Материалы международной конференции. Актобе, 2019, 18-24 с.

49. Сиразитдинов З.Ә. Қазақстандағы қолданбалы лингвистика құрылысында Жұбановтардың рөлі туралы// «Жұбанов тағылымы» атты X халықаралық ғылыми-практикалық конференция материалдары. – Ақтөбе: Қ.Жұбанов атындағы АӨМУ, 2019. 28-36 б.

50. Сиразитдинов З.А. О принципах создания электронного информационного ресурса произведений русскоязычных писателей Республики Башкортостан//«Жұбанов тағылымы» атты X халықаралық ғылыми-практикалық конференция материалдары. – Ақтөбе: Қ.Жұбанов атындағы АӨМУ, 2019. 226-231 б.

51. Бускунбаева Л.А., Сиразитдинов З.А. Принципы транскрибирования аудиоматериалов в диалектном корпусе башкирского языка// Финно-угорскийй мир в полиэтничном пространстве России: культурное наследие и новые вызовы. Ижевск, 2019. С. 545-548.

52. Сиразитдинов З.А. О применении корпусов в практике русско-башкирского перевода// Финно-угорский мир в полиэтничном пространстве России: культурное наследие и новые вызовы. Ижевск, 2019. С. 554-561.

53. Бускунбаева Л.А., Сиразитдинов З.А. О принципах транскрибирования аудиоматериалов для фонетических корпусов башкирского языка // Материалы международной научной конференции Turklsng 2019, Симферополь, 2019

54. Сиразитдинов З.А. О концепции создания Национального корпуса башкирского языка//Электронная письменность народов Российской Федерации; опыт, проблемы и перспективы: Материалы II Международной научной конференции. Уфа, 2019.

55. Сиразитдинов З.А. О вкладе Жубановых в становлении прикладной лингвистики в Казахстане//Личности великой степи: Учение Ж убановых и реформа письма: Материалы международной конференции посвященной 120-летию видного общественного деятеля, основателя теории казахского языкознания, первого профессора, педагога-ученого Кудайбергена Ж убанова и 90-летию его сына, доктора филологических наук, профессора Есета Жубанова. Алматы, 2019. С.

56. Бускунбаева Л. А. Хезитационная пауза как метакоммуникационный маркер в башкирском языке (на материале диалектного корпуса башкирского языка) // Международный симпозиум Башкирской культуры, истории и литературы им.М.Акмуллы в рамках Международного саммита Науруз. Анталья, 2019. С. 70–74.

57. Бускунбаева Л. А. Роль хезитационных пауз в башкирской диалектной речи (на материале диалектного корпуса башкирского языка)" // Материалы конференции "Алтайское языкознание: проблемы и исследования" (18-22 июня 2019). Горно-Алтайск, 2019. С. 148–154.

58. Бускунбаева Л. А. Принципы сбора полевых материалов для диалектного корпуса башкирского языка // Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы (11–12 декабря 2019 г.). Уфа, 2019. – С. 3–6.

59. Шамсутдинова Г.Г. «Отражение народных обычаев и обрядов в башкирском народном эпосе «Кузыйкурпес и Маянхылыу» // Материалы

Третьего Съезда молодых востоковедов России и стран СНГ. Уфа, 25-27 сентября 2019 г. – М.: Пробел-2000, 2019. – С.111-112.

60. Шәмсетдинова Г.Ғ. Башҡорт халыҡ эпосы «№уҙыйкүрпәс менән Маянһылыу»ҙың 1973 йылда Белорет районы Мәхмүт ауылында яҙып алынған бер варианты тураһында (башҡорт теленең фольклор корпусы материалдары нигеҙендә) // Башкирский язык и литература в условиях глобализации и полиэтнической среды: опыт и перспективы - Уфа, 2019. С. 405-408.

61. Шамсутдинова Г.Г. «Сказочные мотивы в башкирском народном эпосе «Кузыйкурпес и Маянхылыу» (на материале Корпуса фольклора башкирского языка) // Материалы международной конференции «Түркі жазуының негіздері». – Центр развития языков и перехода на латинскую графику акимата города Алматы. Алматы, 2019. С.44-55.

62. Shamsutdinova G.G. Başkurt halk destani ‘Kuzukürpes ile Mayanhiliu» // Сборник тезисов международной конференции ASEAD VI. ULUSLARARASI SOSYAL BİLİMLER SEMPOZYUMU, 2-4 Kasım 2019 ANTALYA, TURKEY. Турция, Анталья, 2019. С.3.

63. Бускунбаева Л.А., Ишмухаметова А.Ш., Шамсутдинова Г.Г. Машинный фонд как средство обучения башкирскому языку // Материалы II Международной научной конференции Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы – Уфа, 2019. С.

64. Ишмухаметова А.Ш., Шамсутдинова Г.Г. О лексикографической базе данных Машинного фонда башкирского языка // Материалы II Международной научной конференции Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы – Уфа, 2019. С.

65. Бускунбаева Л.А., Ишмухаметова А.Ш., Шамсутдинова Г.Г., Каримова Р.Н. Ученый-новатор (к 60-летнему юбилею известного башкирского лингвиста З. А. Сиразитдинова) // Российская тюркология (в печати).

66. Ишмөхәмәтова А. Ш. “Урал-батыр” эпосында теркәүестәрҙең сағылышы» // Сборник докладов Международного симпозиума Башкирской

культуры, истории и литературы им. М.Акмуллы в рамках международного саммита Науруз. Анталья, 2019 (март). – С. 64-69.

67. Ишмөхәмәтова А.Ш. «Урал-батыр» эпосында үҙ ләштерелгән һүҙ ҙәр (башҡ орт теленең фольклор корпусы нигеҙ ендә тикшереү) // «Башкирский язык и литература в условиях глобализации и полиэтнической среды: опыт и перспективы»: Материалы Международной научно-практической конференции (г. Уфа, 22-25 октября 2019 г.) – Уфа, 2019. – 167-171-се бб.

68. Сиразитдинов З. Ә., Боҫ ҡонбаева Л. А., Ишмөхәмәтова А Ш., Шәмсетдинова Г. Ғ. Башҡ орт теленең диалект аудиокорпусын төҙ өү: проблемалар һәм перспективалар // Ватандаш, №12, Өфө, – 93-102-се бб.

69. Ишмухаметова А.Ш., Шамсутдинова Г.Г. О лексикографической базе данных Машинного фонда башкирского языка // Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы: II Материалы Международной научной конференции (г. Уфа, 11-12 декабря 2019 г.) – Уфа, 2019, С. 64-66.

70. Бускунбаева Л.А., Ишмухаметова А.Ш., Шамсутдинова Г.Г. Машинный фонд как средство обучения башкирскому языку // Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы: Материалы II Международной научной конференции (г. Уфа, 11-12 декабря 2019 г.) – Уфа, 2019, С. 104-106.