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
«28» января 2019 г.

УТВЕРЖДАЮ
Врио председателя УФИЦ РАН
д.э.н.
_____ Р.Р. Ахунов
«28» января 2019 г.

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

СОЗДАНИЕ КОРПУСА ДИАЛЕКТНЫХ ТЕКСТОВ БАШКИРСКОГО ЯЗЫКА

(промежуточный, этап 2)

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

Уфа 2019

ОТЧЕТ

О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

СОЗДАНИЕ КОРПУСА ДИАЛЕКТНЫХ ТЕКСТОВ БАШКИРСКОГО ЯЗЫКА

(промежуточный, этап 2)

Руководитель НИР,

к.ф.н., с.н.с. _____З.А. Сиразитдинов

Уфа 2019

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель темы:
кандидат филологических наук З.А.Сиразитдинов

Исполнители темы:

1. кандидат филологических наук З.А.Сиразитдинов,
2. кандидат филологических наук Л.А.Бускунбаева,
3. младший научный сотрудник Р.Н.Каримова, 0,5 ставки,
4. младший научный сотрудник А.Ш.Ишмухаметова, 0,25 ставки,
5. лаборант-исследователь Г.Г.Шамсутдинова, 0,25 ставки.

РЕФЕРАТ

Отчет 23 с., 42 источников, рис.5.

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.

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 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 … … … … … … … … … … … … . … … … … … … … … … … . 5

2. ОСНОВНАЯ ЧАСТЬ… … … … … … … … … … … … . … … … … … … . 7

2.1. СБОР ПОЛЕВОГО МАТЕРИАЛА… … … … … … … … … … … … . … . 7

2.2. ПРИНЦИПЫ ТРАНСКРИБИРОВАНИЯ… … … … … … … … … 11

2.3. СРЕДСТВА РАБОТЫ С КОРПУСОМ АУДИО И ТЕКСТОВЫХ ДАННЫХ ……. 17

2.4. ПРОВЕДЕННАЯ РАБОТА … … … … … … … . … … … … … … … . 19

3. ЗАКЛЮЧЕНИЕ… … … … … … … . … … … … … … … … … … … … 21

4. ИСТОЧНИКИ И ЛИТЕРАТУРА … … … … … … … … … … … … … 22

1. 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 в качестве самостоятельной [1-11]. В России известны **Саратовский диалектологический корпус**, Мультимедийный корпус диалектных текстов Устьянского района Архангельской области, диалектные подкорпусы в составе Национального корпуса русского языка, ижемского говора коми языка, Национального корпуса калмыцкого языка и др [12-17].

Следует 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 [18].

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 [19]. 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].

2. 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.

.

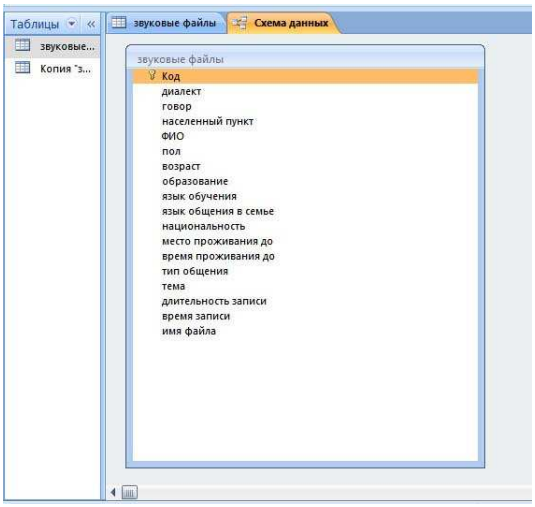

Рис. 1. Схема данных паспортизации файлов и экстралингвистической разметки в базе Access.

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 (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, h 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 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) 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 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 случах, 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) 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 ее, 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, 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, +ҡайтыб*ара – ҡайтып 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.

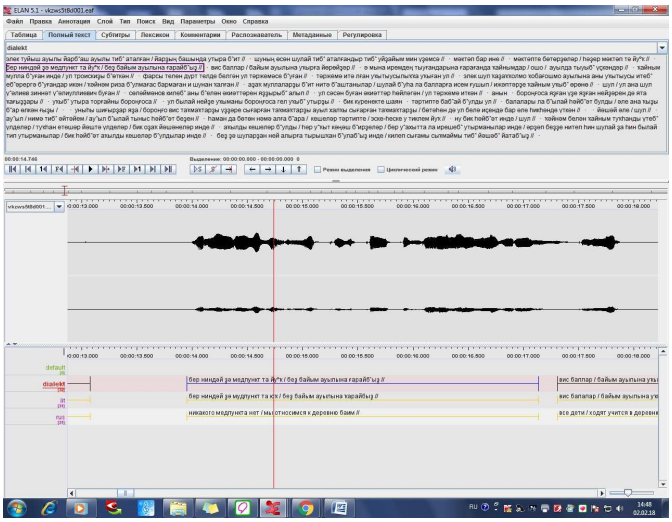

Рис. 2. Скриншот транскрибирования диалектного материала на ELAN.

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

Собранный и аннотированный текстовый материал предполагается внедрить в Машинный фонд башкирского языка и в международный проект Института языкознания РАН [http://lingvodoc.tsu.ru/]. На Рис. 3. представлен скриншот выдачи запроса с учетом всех полей поиска в Машинном фонде башкирского языка.

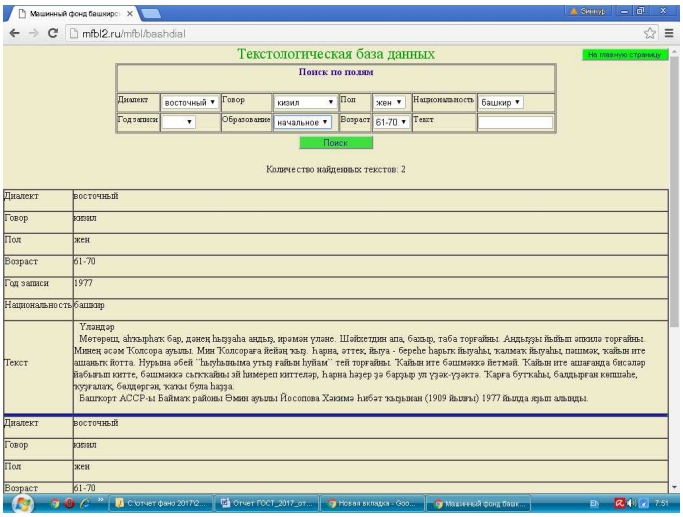

Рис. 3. Пример выдачи текста по сложному запросу.

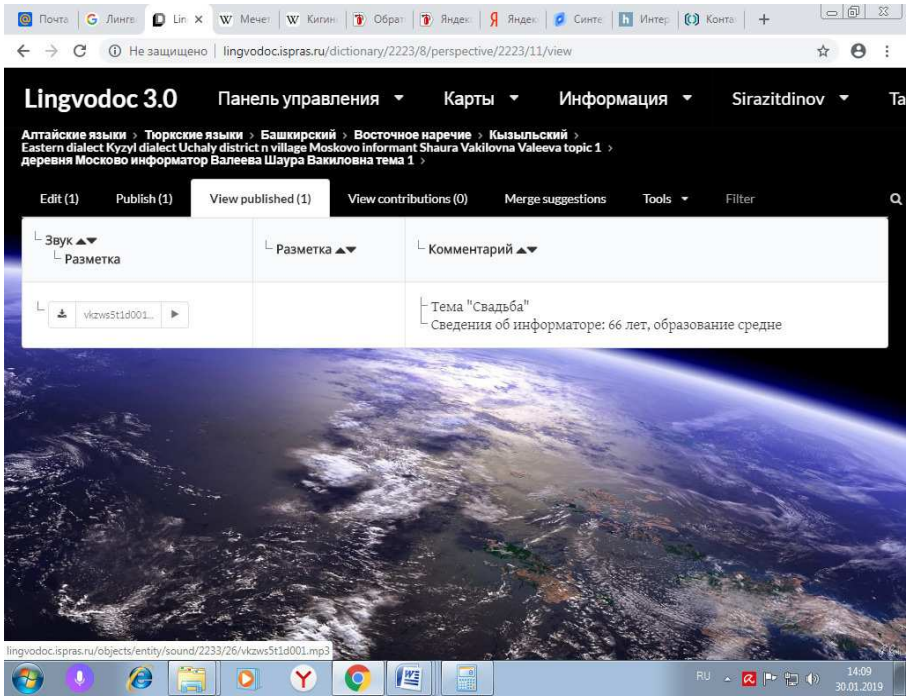

Рис. Скрин-шот представления записи по теме “Свадьба” информатора Шауры Вакиловны Валеевой в Лингводоке.

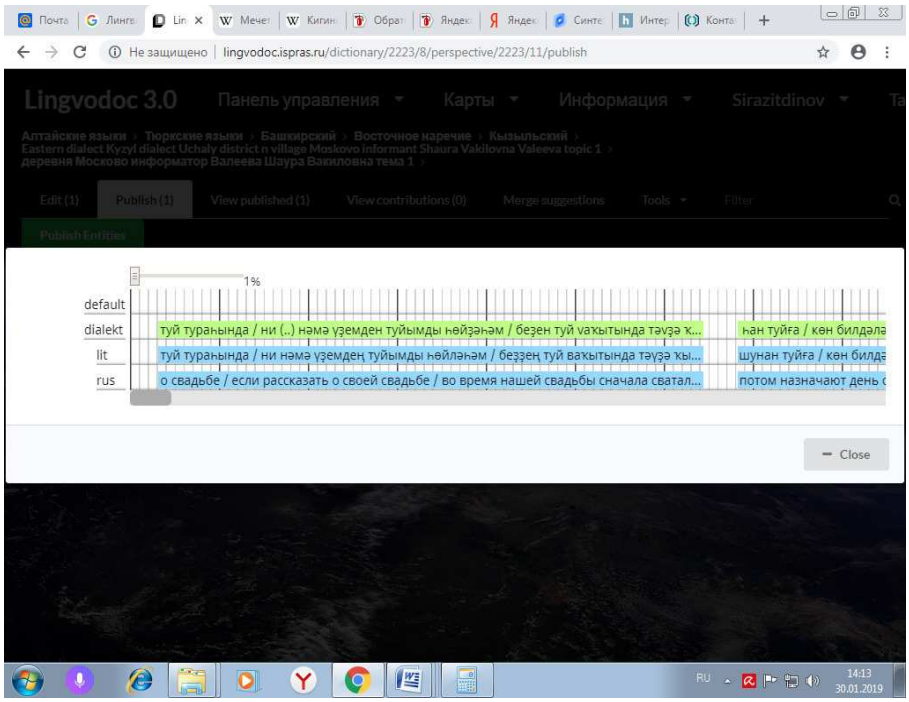

Рис. 5 Рис. Скрин-шот выдачи транскрипции диалектного

и литературного варианта речи и русского перевода по теме “Свадьба”

в Лингводоке

2.4.ПРОВЕДЕННАЯ РАБОТА

1) Проведен сбор полевых материалов по учалинскому говору (Учалинский район РБ, с. Кунакбаево, д. Ильтабаново, д. Юлдашево) и айскому говору (Мечетлинский: с. Таишево, с. Бургаджинов, с. Лемез-Тамак, д. Абрахманово; Кигинский районы РБ: с. Абзаево, д. Ибраевво) восточного диалекта башкирского языка Всего собран материал от 20 информаторов, представляющих 190 записей по заданным темам объемом в 1,7 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) Осуществлено добавление слоев аннотирования «литературный» и «русский», в которых даны литературные варианты транскрипции речи и его перевод на русский язык собранных материалов в 2017 и 2018 гг.

8) Начат ввод транскрибированных материалов из формата ELAN в текстовую поисковую систему Машинный фонд башкирского языка (mfbl2.ru).

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

3. ЗАКЛЮЧЕНИЕ

За время работы в 2018 г. доработаны принципы подготовки и обработки полевого материала. Обработан речевой материал экспедиций по двум говорам: учалинский и айлинский. Начато представление материалов в информационных базах данных “Машинный фонд башкирского языка” и “Лингводок”.

В рамках научного проекта коллективом опубликовано 14 научных статей [29-42] (из них 5 в рецензируемых журналах).

В 2019 г. предстоит:

- провести сбор новых полевых речевых материалов по миасскому, аргаяшскому и сальютскому говорам;

- провести первичную обработку полевого материала;

- осуществить сегментацию речевого материала;

- транскрибировать и аннотировать речевой материал;

- закончить ввод всех материалов в базы данных “Машинного фонда башкирского языка” в раздел “Диалектологическая база данных” и в базу данных “Лингводок”.

4. ИСТОЧНИКИ ЛИТЕРАТУРА

1. **Das Deutsche Referenzkorpus** (DeReKo) URL:**http://www.ids-mannheim.de/kl/projekte/korpora** (дата обращения: 25.12.2017)**.**

2. British National Corpus (BNC). URL:**http://www.natcorp.ox.ac.uk** (дата обращения: 25.12.2017)**.**

3. The corpus of contemporary american english (COCA). URL:**http://corpus.byu.edu/coca** (дата обращения: 25.12.2017).

4. Корпус китайского языка. (LIVAC Synchronous Corpus). URL:**http://www.rcl.cityu.edu.hk/livac** (дата обращения: 25.12.2017)**.**

5. Magyar Nemzeti Szövegtár (корпус венгерского языка). URL:**http://corpus.nytud.hu/mnsz** (дата обращения: 25.12.2017).

6. Corpus del español (корпус испанского языка). URL:**http://www.corpusdelespanol.org** (дата обращения: 25.12.2017)**.**

7 LDC Top Ten Corpora (мультимедийные корпусы английского языка). URL: http://www.ldc.upenn.edu/Catalog/topten.jsp (дата обращения: 17.12.2017).

8. Фонетический немецкого разговорного языка URL:http://dsav-wiss.ids-mannheim.de/korpora/pf/pf _doku.htm (дата обращения: 25.01.2017).

9. Chinese Broadcast Conversation Speech (мультимедийный корпус китайского языка). URL: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId= LDC2013S04 (дата обращения: 17.12.2017).

10. 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 25.12.2017).

11. ***The Spoken Turkish Corpus (***мультимедийный корпус разговорного турецкого языка). URL:http://stc.org.tr (дата обращения: 25.12.2017).

12. Фонетический корпус “Один день русского языка” URL:http://www.model.org.spbu.ru (дата обращения: 17.01.2018, доступ ограничен).

13. Корпус ижемского диалекта коми языка URL:http://videocorpora.ru/ru/allmedia (дата обращения: 25.12.2017).

14.Саратовский мультимедийный диалектный корпус. ruscorpora.ru/sbornik2005/13letuchy.pdf. (дата обращения: 25.01.2018).

15. Мультимедийный корпус диалектных текстов Устьянского района Архангельской области URL: http:http://www.slavist.de/Pushkino/ (дата обращения: 25.01.2018).

16. Диалектный подкорпус национального корпуса русского языка URL: http://www.ruscorpora.ru/search-dialect.html (дата обращения: 25.01.2018).

17. Шерстинова Т. Ю., Степанова С. Б., Рыко А. И. Система аннотирования в звуковом корпусе русского языка «Один речевой день» // Материалы XXXVIII международной филологической конференции. Формальные методы анализа русской речи. 16–20 марта 2009 г. Санкт-Петербург. СПб., 2009.

18.Информационная система Лингводок. URL: http:http://lingvodoc.ispras.ru. (дата обращения: 25.01.2019).

19. 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: 25.01.2019).

20. Крючкова 2008 – *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: 20.09.2018).

21. Кибрик и др. 2003 – *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.**

22. ***Степанова С. Б., Асиновский А. С., Богданова Н. В., Русакова М. В., Шерстинова Т. Ю. З*вуковой 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.**

23. 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: 20.09.2018).

24. *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.**

25. 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: 20.09.2018).

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

27. 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.

28. 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.

29. **Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Шамсутдинова Г. Г. О сборе материала для корпуса Башкирской диалектной речи // Материалы Международной научно-методической конференции “Новые научные направления казахского языкознания”, посвященной 90-летию заслуженного деятеля науки РК, доктора филологических наук, профессора Нуржамал Оралбай. – Алматы, 2018. – С. 193–198.**

30. **Сиразитдинов З. А., Бускунбаева Л.А., Ишмухаметова А. Ш., Шамсутдинова Г. Г. Информационные ресурсы в освоении башкирского языка городским населением // Городские башкиры: проблемы сохранения этничности: Материалы X Межрегиональной научно-практической конференции (г.Янаул, 11 мая 2018 г.). – Уфа, 2018. С. 111–116.**

31. Сиразитдинов З. А., Бускунбаева Л. А., Ишмухаметова А. Ш., Шамсутдинова Г. Г. К принципам сбора и обработки полевых материалов по восточному диалекту башкирского языка для диалектного корпуса // I. Uluslararasi Miftahetdin Akmulla Başkurt kültürü, tarihi ve edebiyati sempozyumu. Анталья, 2018. С. 171–179.

32. **Сиразитдинов З. А., Бускунбаева Л.А., Ишмухаметова А. Ш., Шамсутдинова Г. Г.** Проблемы разработки диалектного корпуса башкирского языка // Вопросы диалектологии: Международный научный журнал, № 1–2, 2018. С. 97–107.

33. Сиразитдинов З.А., Ишмухаметова А.Ш. Башҡорт телене\ лингвистик корпустары тураһында// История. Культура: III Юсуповские чтения: Материалы Международной научной конференции, посвященной памяти Рината Мухаметовича Юсупова (1951–2011). 2018. С. 217–220.

34. Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Шамсутдинова Г.Г. О лингвистических корпусах башкирского языка В сборнике: XXI ғасырдың ғылыми-ағартушылық парадигмалары: рухани құндылықтар аясындағы тіл және мәдениет: атты халықаралық ғылыми-практикалық конференция материалдары. Алматы, 2018. С. 276-280.

35. Sirazitdinov Z., Buskunbaeva L., Işmuhametova A., Şamsutdinova G. Başkurt dilinin doğu ağzi derlemi için araştirma seferleri esnasinda materyaller toplama ve işleme prensipleri // II.uluslararasi Miftahetdin Akmulla Baskurt kültürü, tarihi ve edebiyati sempozyumu. Ankara, 2018 (в печати)/

36. Сиразитдинов З.А. О парных сочетаниях именных грамматических аффиксов башкирского языка//Tel – 2018, Казань, 2018.

37. Buskunbaeva L.A., Muratova R.T., Suleymanova R.A. Bashkir Phraseological Units with Numeral Components: Structural and Semantic Aspects // Proceding of 2018 4th EER International Conference on Social Sciences and Interdisciplinary Studies (EER-SSIS-2018). (August 18-20, 2018. Palermo, Sicily, Italy. P. 115-120.

38. Сиразитдинов З.А., Бускунбаева Л.А., Шамсутдинова Г.Г. Типы односоставных предложений в пословичном фонде башкирского языка (на материале фольклорного корпуса)//Вестник КИГИ РАН, №6, 2018. С.120-129.

39.Gulgina Şamsutdinova. “Başkurt destanı Kuzukürpes ile Mayanhılıu’ ınvaryantlarının ve dilinin karşılaştırmalı incelenmesi” (на турецком языке) в сборнике докладов I Международного симпозиума Башкирской культуры, истории и литературы им.М.Акмуллы в рамках международного саммита Науруз, 2018. С.1-7.

40. Хусаинова Г.Р., Сиразитдинов З.А., Шамсутдинова Г.Г. О языке и стиле эпоса «Кузыйкурпяс и Маянхылу» в записи Сайфуллы Салимова // Актуальные проблемы диалектологии языков народов России: МатериалыXVIII Всероссийской научной конференции (с международным участием). Уфа: ИИЯЛ УФИЦ РАН. 2018. С. 128-133.

41. Хусаинова Г.Р., Шамсутдинова Г.Г. “Yeni yazmalarda Başkurtlarin mitolojikhikayeleri” // I. Uluslararasi Develi-aşik seyrani ve türk kültürü kongresi (I.International Develi-asik seyrani and Turkish culture congress), Develi. C. 98-103;

42. Хусаинова Г.Р., Шамсутдинова Г.Г. “Башҡорт һөйләү теленең һәм халыҡ ижадының хәҙерге торошо (Белорет районында яҙып алынған материалдар нигеҙендә)” // II.Uluslararasi Miftahetdin Akmulla Baskurt Kültürü, Tarihi ve Edebiyati Sempozyumu. Ankara, 2018. C.7-8