УДК 811.512.145+ 811.161.1'373

Гибадулин Рустем Яхьевич, директор института перспективных исследований, ФБГОУ ВО «Московский педагогический государственный университет», 109240, Москва, ул. Верхняя Радищевская, 16-18

e-mail: rya.gibadulin@mpgu.edu

Гибадулин Яхья Набиуллович, генеральный директор АНО Информационно-издательский центр «Инсан», 119019, Москва, Гоголевский б-р, 6, стр. 1 (а/я 214)

e-mail: insan6@hotmail.com

Селендили Лемара Сергеевна, доктор филологических наук, профессор кафедры крымскотатарской филологии Таврической академии (СП) ФГАОУ ВО «Крымский федеральный университет им. В.И. Вернадского», Республика Крым, 95007, г. Симферополь, ул. Беспалова, 45-б, к. 214

e-mail: lemara2002@hotmail.com

СПЕЦИАЛИЗИРОВАННАЯ ПРОГРАММА ДЛЯ СОЗДАНИЯ ЛИНГВИСТИЧЕСКИХ ЗВУКОВЫХ АРХИВОВ КРЫМСКОТАТАРСКОЙ РЕЧИ “TATRECORDER”. НАЗНАЧЕНИЕ, СТРУКТУРА, ОПИСАНИЕ[1]

Аннотация: в статье рассматривается описание специализированной программы для записи лингвистических звуковых архивов крымскотатарской речи “TatRecorder”, показан мультимедийный продукт, использующий создаваемый звуковой архив для решения научных и образовательных задач, а именно, озвученный электронный словарь крымскотатарских диалектов. Данный мультимедийный продукт позволит собирать, структурировать и обрабатывать речевой материал для крымскотатарского языка по единой схеме.

[1] Исследование выполнено при финансовой поддержке Российского гуманитарного научного фонда в рамках проекта 15-34-10116 «Электронный словарь диалектов крымскотатарского языка».

Описывается структура программы записи звука и основные стандарты представления данных.

Ключевые слова: тюркология, крымскотатарский язык, электронный словарь, электронная лексикография, программа звукозаписи, программа TatRecorder

GibadulinRustemYakhievich, Director of the Institute for Advanced Studies, FBGOU VO "Moscow Pedagogical State University", 109240, Moscow, ul. VerkhnayaRadischevskaya, 16-18

e-mail: rya.gibadulin@mpgu.edu

GibadulinYakhyaNabiullovich, General Director, ANO Information and Publishing Center "Insan", 119019, Moscow, Gogolevsky Blvd, 6, p. 1 (PO Box 214)

e-mail: insan6@hotmail.com

SelendiliLemaraSergeevna, Doctor of Philology, Professor of the Crimean Tatar Philology Department of the Taurida Academy (JV) FGAOU VO "Crimean Federal University. IN AND. Vernadsky, Republic of Crimea, 95007, Simferopol, ul.Bespalova, 45-b, room 214

e-mail: lemara2002@hotmail.com

SPECIALIZED PROGRAM FOR CREATING LINGUISTIC SOUND ARCHIVES OF THE CRIMEAN TATAR SPEECH "TATRECORDER". DESIGNATION, STRUCTURE, DESCRIPTION

Abstract: the article describes the description of a specialized program for recording linguistic sound archives of the Crimean Tatar speech "TatRecorder". It shows a multimedia product using the created sound archive for solving scientific and educational problems, namely, an electronic dictionary of Crimean Tatar dialects. This multimedia product will allow to collect, structure and processspeech material for the Crimean Tatar language according to a single scheme. The structure of the

sound recording program and the basic standards for data representation are described.

Keywords: turkology, Crimean Tatar language, electronic dictionary, electronic lexicography, sound recording program, TatRecorder program

В современном мире с его растущей глобализацией языки исчезают с высокой скоростью. В начале этого века было спрогнозировано [1], что через столетие половина из ныне исчезающих языков будут мертвыми. Подобный процесс может вылиться в культурную ассимиляцию, которая обычно приводит к потере подавляемого языка в течение жизни следующих поколений (например, второго поколения иммигрантов) [2].

Исследования крымскотатарского языка до сих пор носят весьма фрагментарный характер. Язык, имеющий некодифицированную литературную норму в связи с отсутствием научного лексикографического описания, в 2010 году уже занесен в Атлас исчезающих языков ЮНЕСКО.

Практика показывает, что быстрое распространение во все сферы жизни цифровых технологий приводит к особенно быстрому вытеснению тех языков, которые не имеют соответствующих современным требованиям программных средств. В этой связи развитие корпусных технологий может послужить не только средством по фиксации современного состояния языков малых народов, но определенным фактором, замедляющим негативные тенденции.

В ходе выполнения проекта (РГНФ) «Электронный словарь крымскотатарских диалектов», была реализована система лексико-семантического и синтморфологического маркирования единиц крымскотатарской лексики[3], создана специализированная программа звукозаписи крымскотатарской лексики “TatRecorder”, и прототип электронного словаря крымскотатарских диалектов (https://tatar-tele.info/html/).

Уже внедренная лексическая база - лексика северного (степного) диалекта крымскотатарского языка - извлечена с помощью программы Wordsmith из текстов произведений Мемета Нузета, написанных на северном (степном)

диалекте крымскотатарского языка). С помощью разработанной программы звукозаписи было произведено пробное озвучивание слов на “А”, ”Б”, “В”, “Г” словника, составленного на базе текстов произведений Мемета Нузета (всего озвучено около 2500 слов).

Описываемая в статье программа записи звука “TatRecorder”, необходима для дальнейшего развития оригинального «Электронного словаря крымскотатарских диалектов», который в дальнейшем может послужить в целях сохранения исчезающего крымскотатарского языка базой фиксации звучащей речи (в качестве дикторов будут привлечены носители крымскотатарского языка).

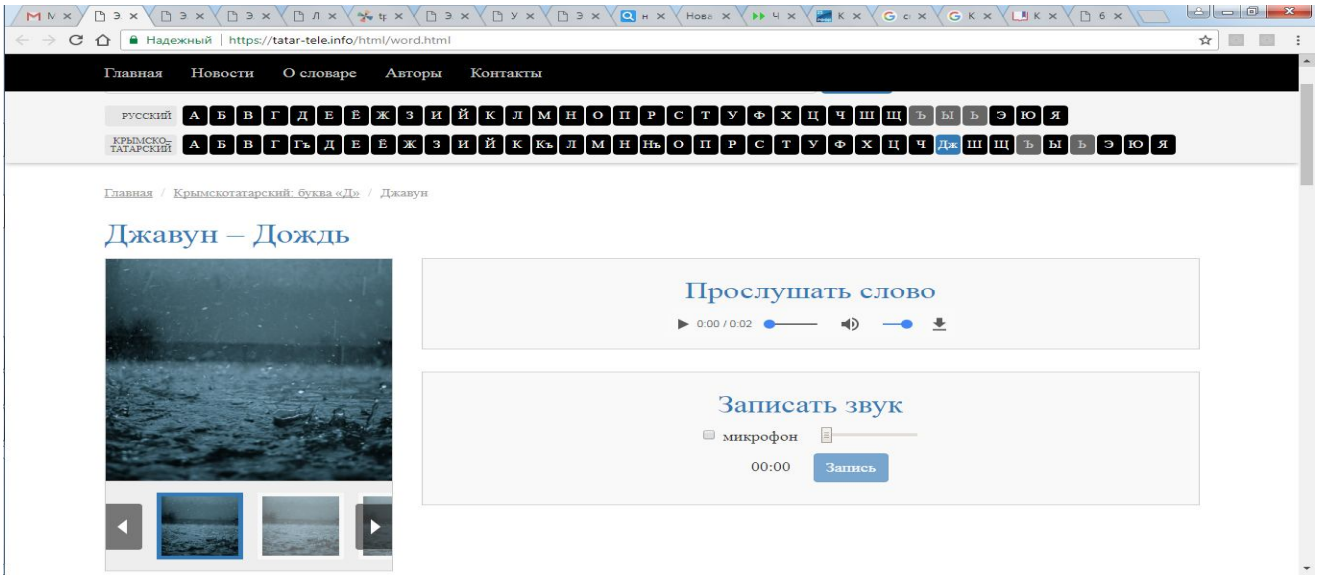

Рис.1. Вид фрагмента словарной статьи «Джавун» с использованием программы звукозаписи (прототип электронного словаря крымскотатарских диалектов (https://tatar-tele.info/html/)

Программа “TatRecorder” разработана как специальная программа для записи и файлирования речевого материала. Программа использует так называемый метод суфлера (prompt-method), который позволяет создавать звуковые файлы, соответствующие отдельным объектам речевого корпуса, непосредственно в процессе его записи, а также создавать аннотацию к оцифрованному фрагменту речевого сигнала (сопроводительную ассоциированную информацию определённого типа).

Разработанная программа будет использоваться и для полевых исследований по данной теме.

Разработка описываемого программного комплекса основана на предыдущих разработках программного обеспечения по татарскому и крымскотатарскому языкознанию, выполняемых в с 2010 года АНО ИИЦ “Инсан” (Гибадулин Я., Гибадулин Р.), а с 2014г. - в сотрудничестве с профессором Селендили Л. С., имеющей опыт разработки электронного «Русско-крымскотатарского словаря лингвистических соответствий» (2005-2010гг.)[4-7].

Описание программы звукозаписи «TatRecorder» /
АНО ИИЦ “Инсан”, 2014-2017

1. Назначение.

Программа TatRecorder разработана для озвучивания текстов (слов, предложений). Может быть использована как рабочий инструмент при создании звуковых приложений для учебных целей, словарей и т.д. Удобна для применения в корпусных технологиях обучения, создания речевых баз данных. Характеристики записи: студийная и полевая запись непосредственно на переносной компьютер в формате wav . Формат звуковых данных: 1) wav PCM (44000 Гц, 16-бит, моно); 2) протоколы сеансов (.ptk файлы).

2. Входные/выходные данные

Входные данные задаются в виде простого неформатированного текстового (plain) файла. Каждый его элемент (слово, предложение) размещается на отдельной строке листинга. Кодировка символов текстов двухбайтовая (UNICODE). Имя файла входных данных – ENTRIESlist (расширение .txt). В файловой системе программы этот файл размещается в каталоге Records. В этом же каталоге размещаются выходные данные: звуковые (.wav)_файлы и протоколы сеансов (.ptk файлы). Имена ENTRIESlist и

Records в программе присвоены соответствующим файлам данных и используются по умолчанию.

3. Главное страница программы и работа диктора. Главная страница дает картину размещения/layout/ управляющих клавиш, полей настройки, размещения поля озвучиваемого текста и др. Картина меняется в зависимости от функционального этапа озвучивания. Всегда содержит листинг слов в нижней части страницы. Часть слов в нем может быть уже озвучена. Каждый сеанс начинается с поиска первого неозвученного слова – входной точки листинга для очередного сеанса. По команде «Найти неозвученное» программа производит поиск такого слова и помещает его в окно/поле/ текстов для озвучивания. В дальнейшем, при озвучивании второго, третьего и т.д. слов в текущем сеансе, программа автоматически выбирает очередное неозвученное слово и помещает соответствующий текст в окно озвучивания. Диктор, озвучивая текст, продолжает сеанс до некоторого момента, например, до перерыва для отдыха.

При приобретении опыта работы с программой, диктор может работать в полуавтоматическом режиме. Программа «снимает» с диктора заботы по программно-техническому обеспечению процессов звукозаписи, включая контроль и документирование. В полуавтоматическом режиме для каждого слова диктор произносит голосом очередной всплывающий в окне текст, нажимает мышкой на одно и то же место на гл. странице (место клавиши “ЗАПИСЬ” и “СТОП).

4. Порядок работы с программой:

4.1 Сеансы озвучивания. Под «озвучиванием» в тексте настоящего руководства понимается произнесение диктором заданного в окне озвучивания текста с одновременной записью его звучания.

Озвучивание выполняется сеансами. Сеанс – это последовательное озвучивание (с возможными перерывами) некоторого количества текстов слов

(предложений). Объем озвучиваемых текстов в сеансе не нормируется и определяется только диктором.

Сеанс запускается по команде «Найти неозвученное» и завершается по команде «Завершить сеанс». Запись каждого очередного слова в сеансе начинается при нажатии клавиши. «ЗАПИСЬ» и прекращается при нажатии клавиши «СТОП».

Для снятия усталости диктора при продолжительной работе разрешены кратковременные перерывы на отдых, длительностью до 20 мин. Диктор, по своему желанию, может сделать перерыв в любое удобное для него время, не прерывая текущий сеанс. Вернувшись до истечения предельного времени, диктор может продолжить сеанс; при этом сведения о перерыве будут внесены в протокол сеанса. Если до истечения перерыва работа не была продолжена, программа автоматически завершает сеанс.

4.2 Исполнение сеансов. Все действия, совершаемые диктором в сеансах, одинаковы. Единственное отличие – первый сеанс, перед которым диктор должен выполнить настройку программы, а именно:

1) в поле справа от слова «Диктор» записать данные идентификации диктора;

2) установить значения параметров оцифровки аналогового сигнала микрофона: частоту выборки(Fr) и разрядность(B). Предустановленные значения: Fr=22050 Кгц, B=16 бит. Эти значения, используемые по умолчанию, оптимальны для оцифровки и качественной звукозаписи большинства речевых сигналов с шириной звукового спектра речи в пределах до 12 кгц. Возможно изменение предустановленных значений. Ввиду особой важности контроля этих параметров, их перенастройка допускается только по паролю администратора. Вход в поле администратора - символ «*» (в верхней части экрана, справа). Диктору рекомендуется использовать предустановленные значения параметров оцифровки;

3) установить предельную длительность (T) интервала озвучивания (по умолчанию T=15 сек);

4) очистить, при необходимости, каталог Records от звукоданных предыдущих сеансов;

5) записать цифру 1 в счетчик номера сеанса, если желательно не продолжать нумерацию сеансов, а начать отсчет с начала.

4.3 Контроль озвучивания возможен на всех этапах звукозаписи. Первичный контроль – прослушивание звучания слова сразу же после окончания произнесения. Групповой контроль - прослушивание сразу же после завершения сеанса. Итоговый контроль - прослушивание звукозаписей, произведенных в различных сеансах, в алфавитном порядке следования слов. Во всех случаях дефектная звукозапись может быть удалена диктором (оператором) и соответствующее слово(текст) автоматически возвращен на повторное озвучивание.

4.4 Протокол сеанса. Каждый сеанс озвучивания документируется. В процессе озвучивания создается протокол сеанса – текстовый (.ptk) файл, содержащий информацию о сеансе (№сеанса, сведения о дикторе, дате и времени озвучивания слов, перерывах и др). Протокол является паспортом звукозаписей в сеансе.

4.5 Звукозапись происходит с момента нажатия клавиши “ЗАПИСЬ” и заканчивается при нажатии клавиши “СТОП”. Клавиша “ЗАПИСЬ” после нажатия становится клавишей “СТОП” и, наоборот, т.е. она совмещает обе функции. Клавиша ЗАПИСЬ/СТОП находится в одном и том же месте главной страницы и, благодаря этому, в процессе звукозаписи мышка всегда оказывается на рабочей клавише и не требует перемещения.

Для качественной записи следует начинать произнесение текста голосом с небольшой задержкой (примерно полсекунды) после нажатия клавиши “ЗАПИСЬ” и нажимать на клавишу “СТОП” (примерно через полсекунды) после окончания произнесения. При ошибках в синхронизации этих моментов, начало либо конец слова не будут озвучены. Предельное время, отведенное на произнесение текста в окне озвучивания, равно 15 сек. При обнаружении

дефектов озвучивания, соответствующая звукозапись удаляется и «плохо» озвученное слово возвращается на повторное озвучивание в следующем сеансе

4.6 Индикатор состояния записи. При появлении текста слова в окне озвучивания индикатор окрашивается в различные цвета в зависимости от состояния записи. Желтый цвет означает – слово не озвучено; зеленый – уже озвучено; красный цвет – дефект при озвучивании - превышение предельного времени записи(текст: «файл плохой – повторить»).

5. Пример работы программы в сеансе озвучивания3-х слов:АБАДАН, АБАЙСЫЗ, АБДЕС. Ниже на рис.2-11 показаны скриншоты главной страницы программы при озвучиваниитрех слов из крымскотатарского словаря М.Нузета.

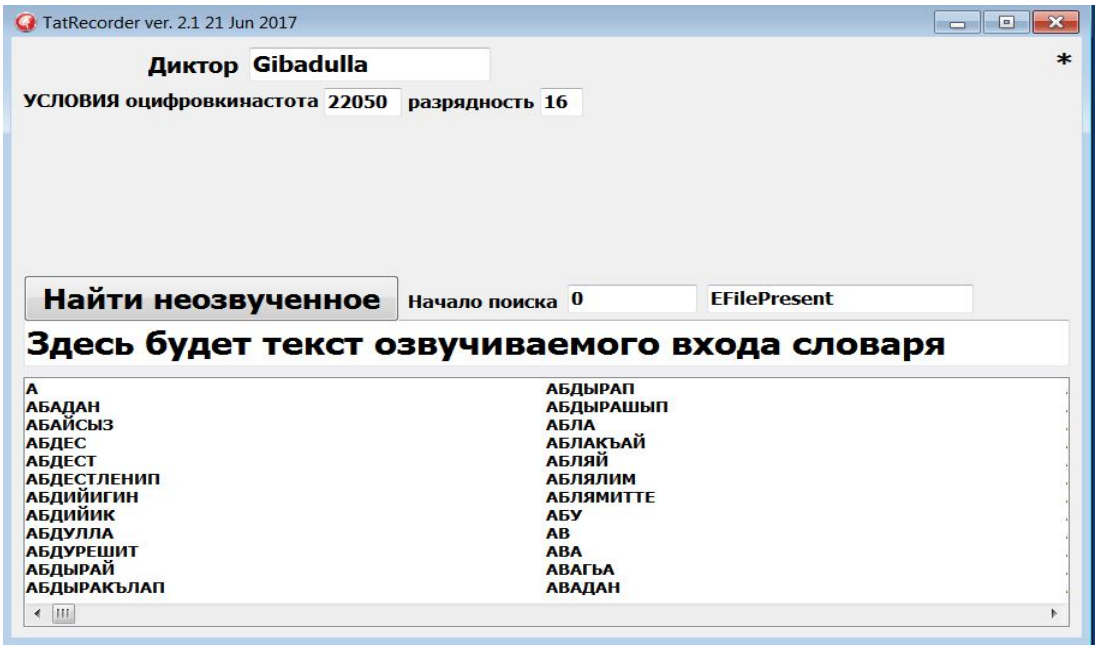

Рис.2. Исходный вид главной страницы программы TatRecorder.

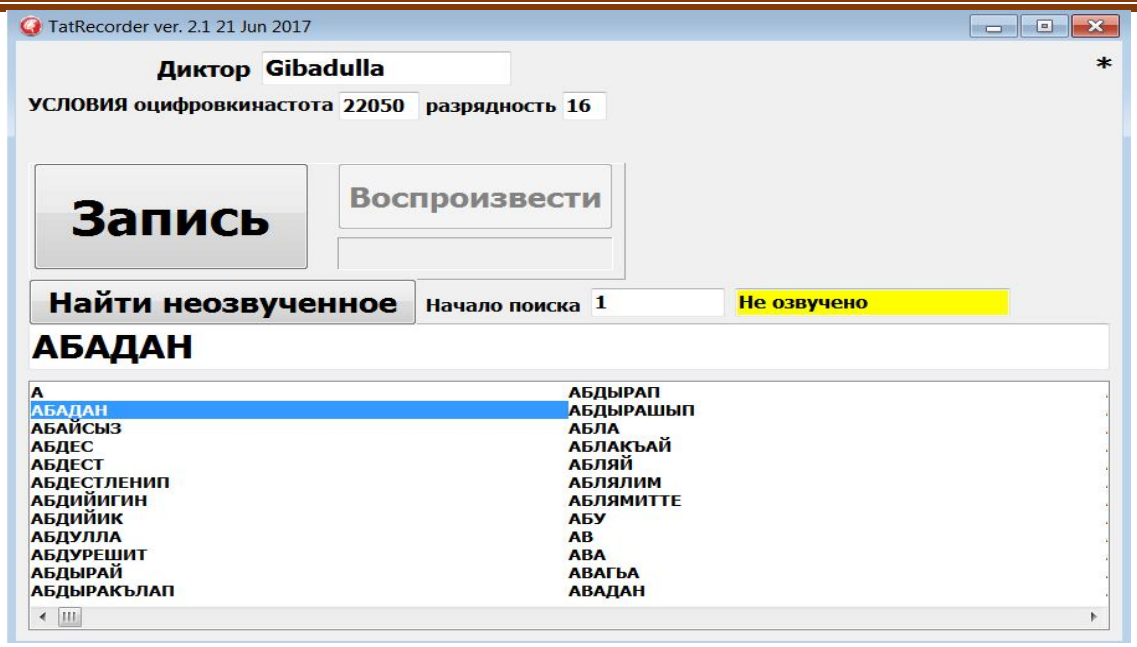

Рис.3. Вид главной страницы после нажатия клавиши «Найти неозвученное».

Первое неозвученное слово АБАДАН помещено в окно озвучивания, индикатор состояния записи окрашен в желтый цвет. Система готова начать озвучивание первого слова. Старт - нажатие клавиши «ЗАПИСЬ». При правильной работе - полсекунды в начале - протяжка, затем, запись голоса диктора при чтении текста (произнесения слова Абадан). Конец записи - при нажатии клавиши «СТОП» (рис.4) через полсекунды после прекращения произнесения слова Абадан.

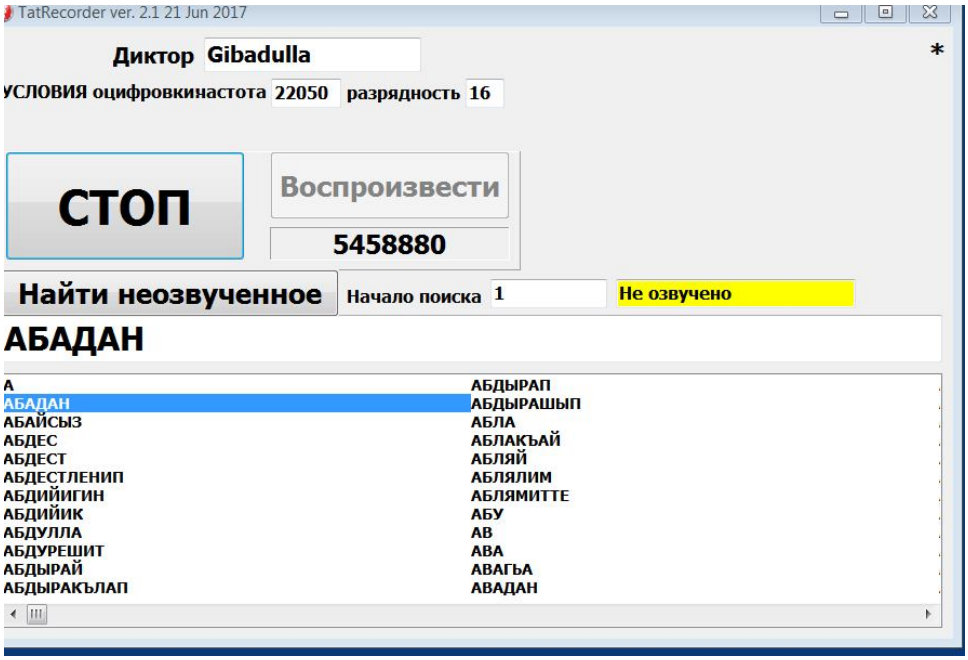

Рис. 4. Вид главной страницы сразу после нажатия клавиши «ЗАПИСЬ». Показан момент начало озвучивания слова АБАДАН.

Озвучивание еще не завершено (желтый цвет индикатора). Объем уже произведенной записи 5458880 байт. Клавиша «ЗАПИСЬ» (рис.4) превращается в клавишу «СТОП»(рис.5).

Для прекращения записи, диктор должен нажать клавишу «СТОП» (рис.5). Вид главной страницы после этого нажатия и завершения озвучивания слова АБАДАН показан на рис.5. В окне озвучивания (рис.6) появилось очередное неозвученное слово АБАЙСЫЗ (цвет индикатора – желтый).

Программа (рис.5) готова к озвучиванию нового слова. Вид главной страницы программы такой же, что на рис.3, но в окне озвучивания слово АБАЙСЫЗ сменило слово АБАДАН и появилась новая клавиша «Завершить сеанс».

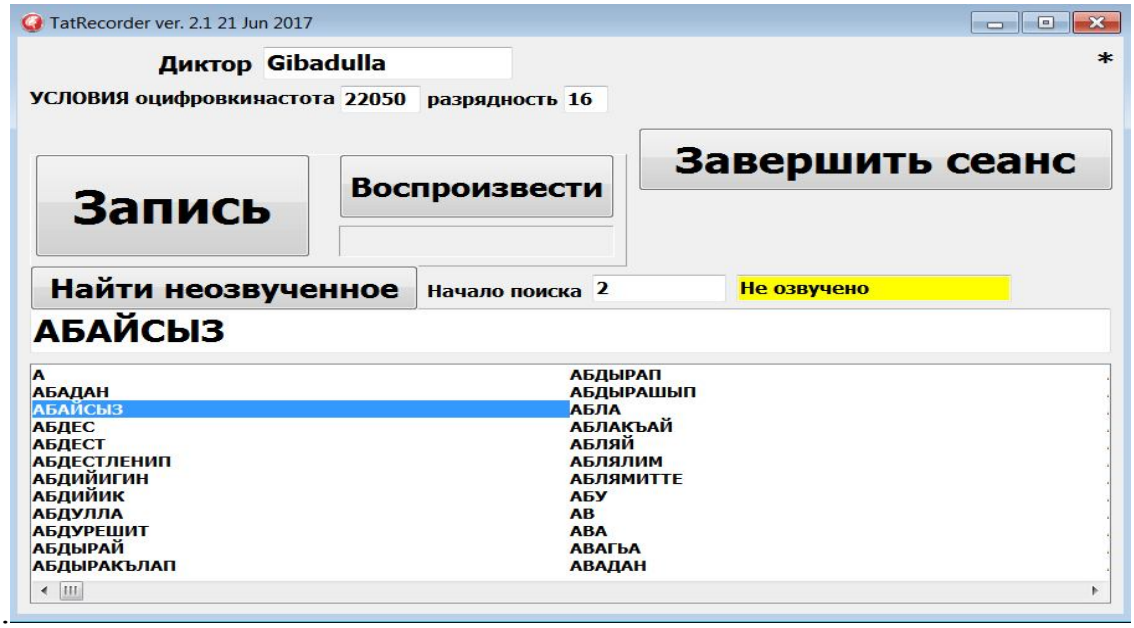

Рис. 5. Вид главной страницы после нажатия «СТОП»и прекращения звукозаписи слова АБАДАН.

В окне озвучивания появилось неозвученное слово АБАЙСЫЗ. Клавиша «СТОП» (рис.6) вновь стала клавишей «ЗАПИСЬ» Появилась клавиша «Завершить сеанс».

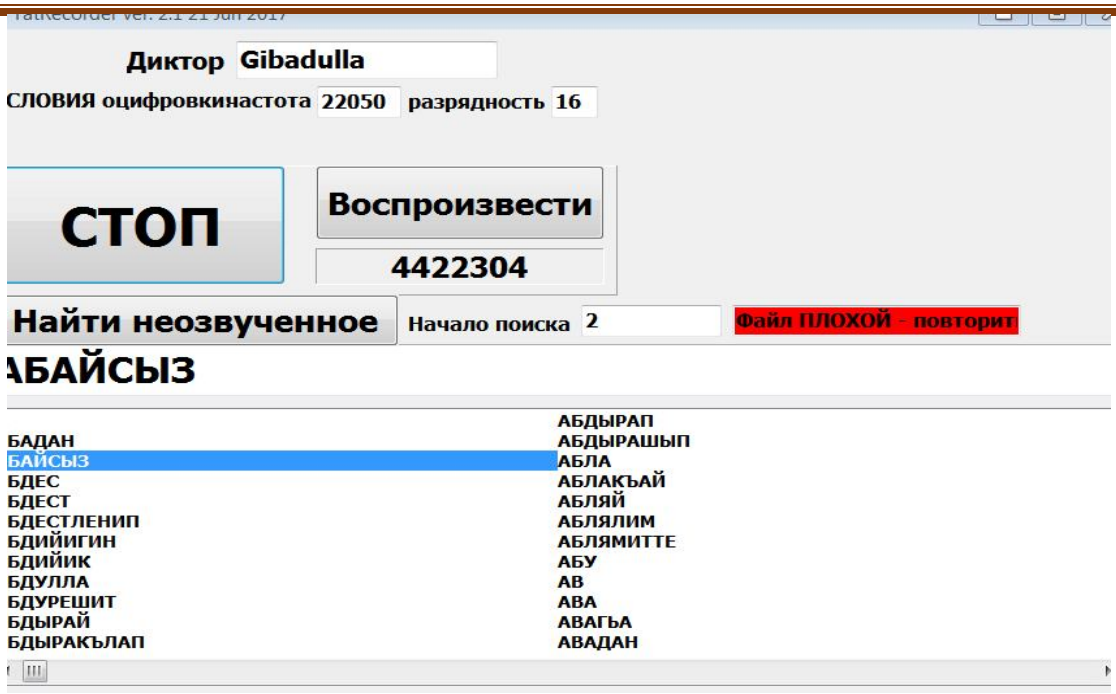

Рис.6. Вид главной страницы после нажатия клавиши «ЗАПИСЬ».

Пример записи с превышением предельной длительности. Клавиша «СТОП» нажата с опозданием. Индикатор состояния записи – красного цвета.

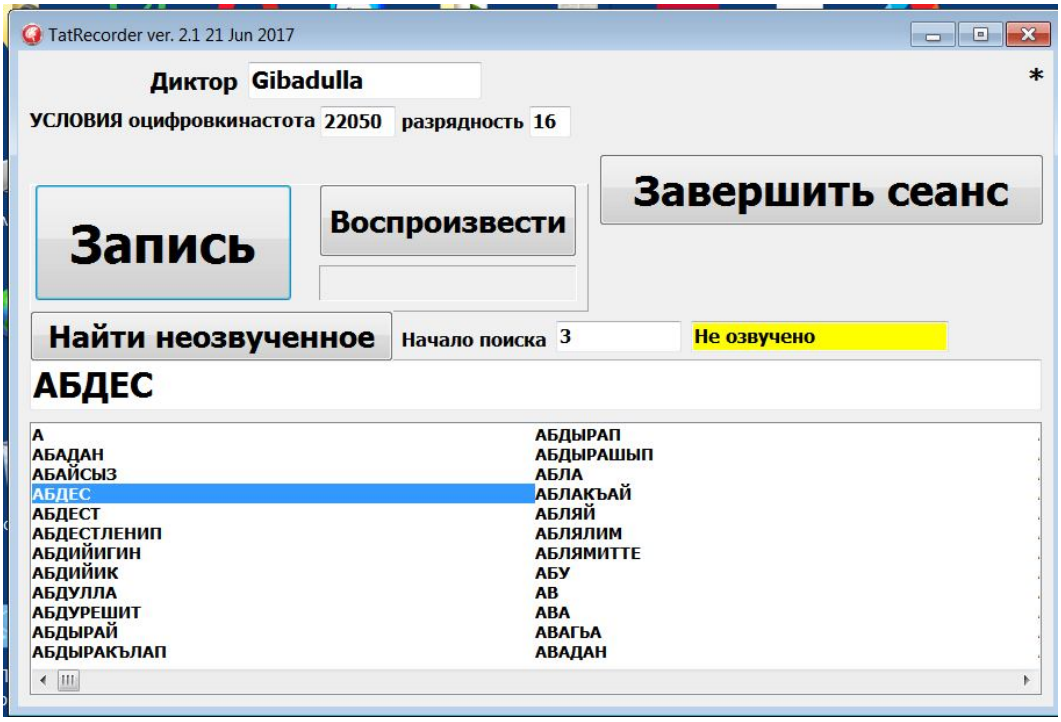

Рис.7. Начальное состояние главной страницы программы при озвучивании слова АБДЕС.

Повтор операций со сменой озвученных слов. Исходное состояние начала озвучивания любого слова (предложения). Далее нажатие клавиши «Завершить сеанс» (рис.8).

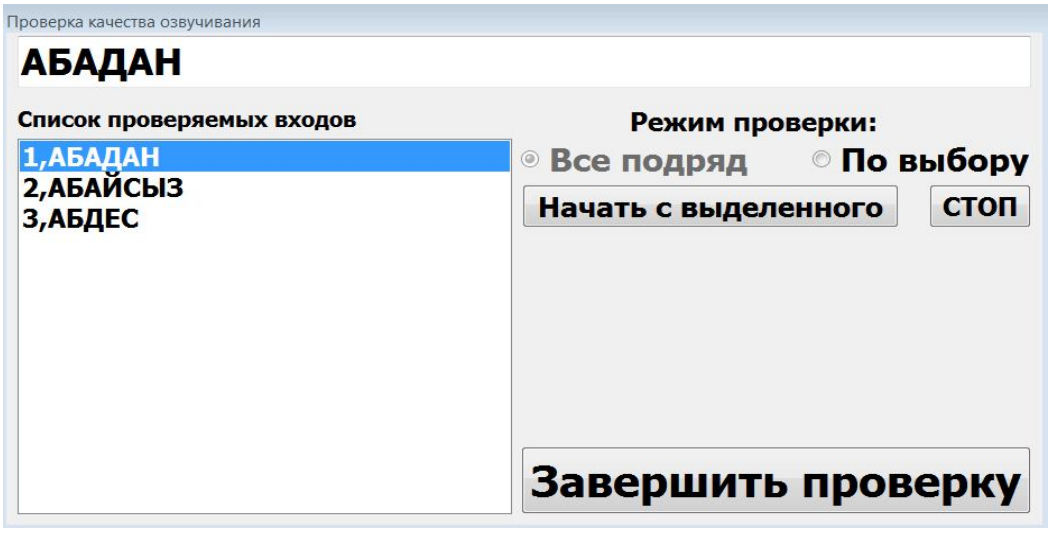

Рис.8. Завершение сеанса. Режим «Все подряд».

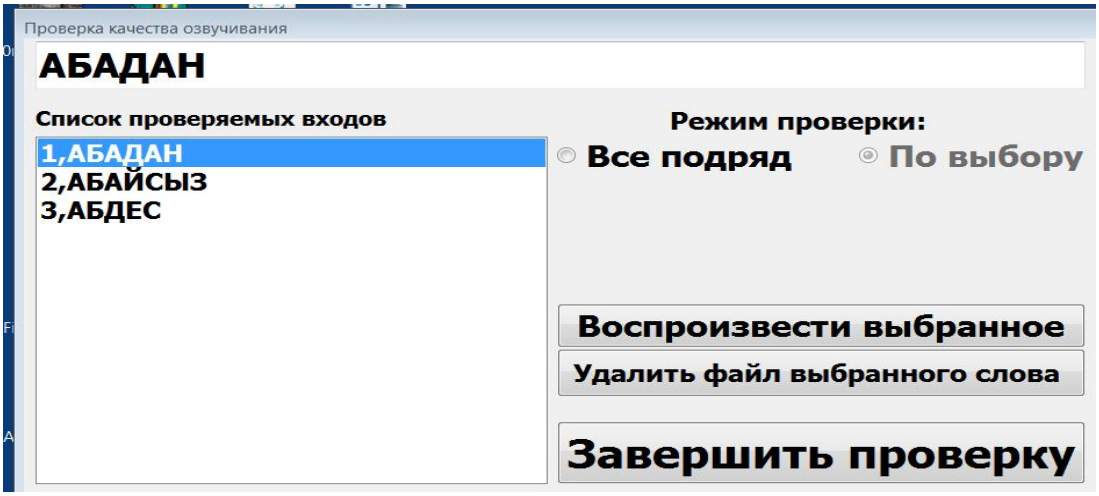

Рис.9. Завершение сеанса. Режим проверки «По выбору».

Проверка качества записи после завершения сеанса.Запускается проверка «подряд», а при обнаружении дефектного звучанияможно перейти на режим «По выбору»; выделить дефектное слово и удалить соответствующую ему звукозапись нажатием клавиши «Удалить файл выбранного слова».

Нажатие этой клавиши возвращает слово в листинг неозвученныхи вынуждает программу произвести повторное озвучивание в следующих сеансах.

Завершение сеанса возможно и без полной проверки озвученного в сеансе с помощью клавиши «Завершить проверку». Нажатие этой клавиши до окончания проверки всего озвученного в сеансе приведет к закрытию сеанса с «хвостом». В этом случае непроверенные «хвосты» перейдут для проверки в следующий сеанс.

ПЕРЕРЫВЫ. Если диктор не взаимодействует с программой в течение одной минуты, то она посылает диктору сообщение о начале отсчета допустимого времени перерыва(20 мин). Формат сообщения показан на рис. 10.

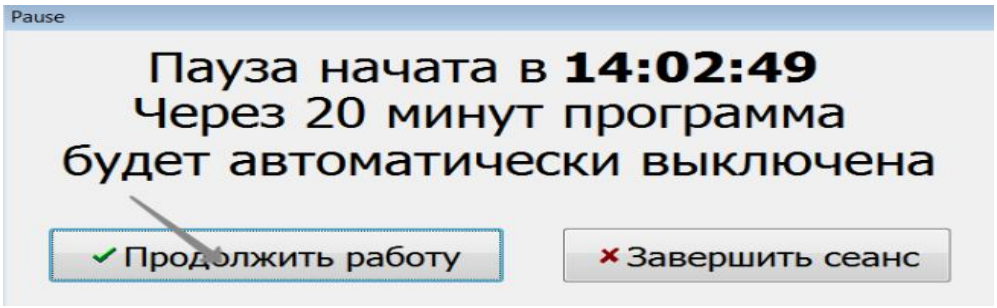

Рис.10. Формат сообщения о начале перерыва.

6. ФОРМАТ ПРОТОКОЛА СЕАНСА

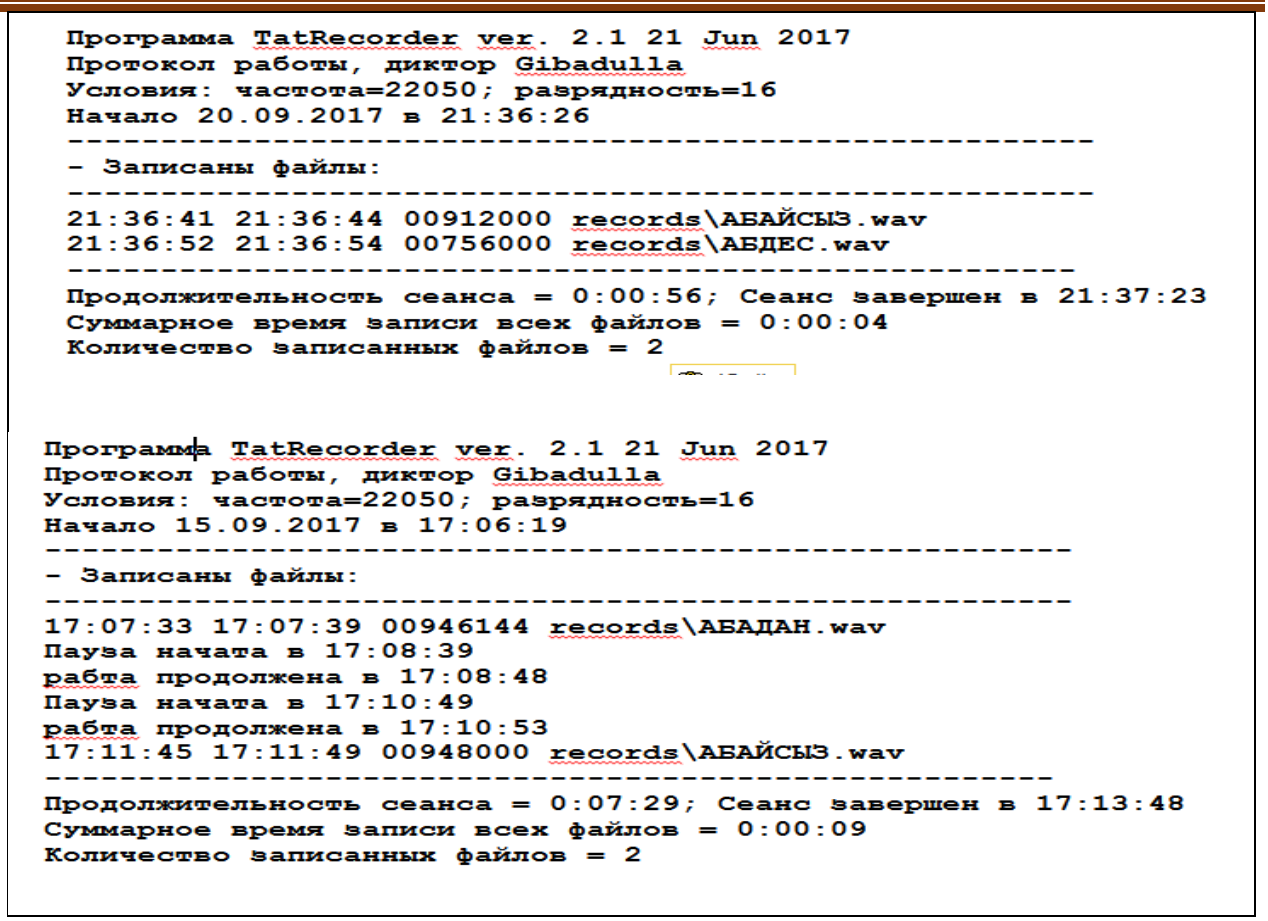

Рис.11. Сеанс №1 Слова: Абадан, Абайсыз. Сеанс №2 слова: Абайсыз, Абдес

7. Итоговая проверка озвучивания. При большом объеме исходных текстов, количество сеансов озвучивания может быть очень большим. Сеансы могут быть записаны в разное время и озвучены разными дикторами, содержать различное количество звукозаписей. С течением времени образуется большой архив звукоданных. В связи с этим часто возникает потребность в систематизации и упорядочении, проведении контрольных прослушиваний, документировании звукозаписей. Для решения такого рода задач создана утилита «SOUNDCONTROL».

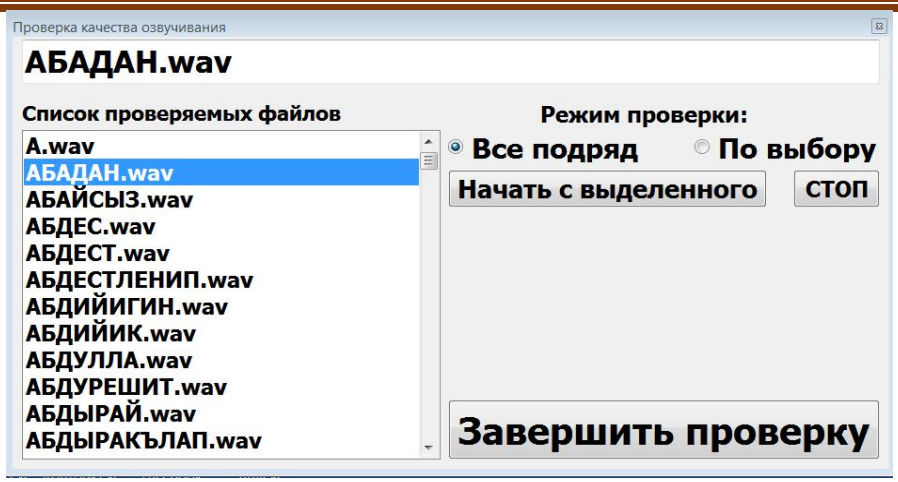

Рис.12 Итоговая контрольная проверка озвученного по всем сеансам.

Разработанная нами в процессе выполнения проекта при финансовой поддержке Российского гуманитарного научного фонда (Российского фонда фундаментальных исследований) № 15-34-10116 «Электронный словарь диалектов крымскотатарского языка» программа звукозаписи позволит проводить полевые исследования по крымскотатарскому языку. В перспективе планируется звуковая фиксация диалектного материала крымскотатарского языка.

Список источников:

1. David Crystal, Language Death, Cambridge University Press, 2000. – P. 198.

2. КарповА.А., Верходанова В.О. Речевые технологии для малоресурсных языков мира [Speech technologies for under-resourced languages of the world] //Вопросыязыкознания [VoprosyJazykoznanija], 2015. – № 2.– С.117-135.

3. Селендили Л.С. Микросинтаксис крымскотатарского языка (формальный и прикладной аспекты). Таврический нац. ун-т имени В. И. Вернадского. – Симферополь: ДИАЙПИ, 2012. – 348 с.

4. Гибадулин Я. Н., Селендили Л. С., Гибадулин Р. Я. Опыт лексикографического описания северного (степного) диалекта крымскотатарского языка (на материале сборника произведений Мемета Нузета «Къырымнынъ чёль аятындан» // Ученые записки Крымского федерального

университета имени В.И. Вернадского. Серия: Филологические науки, 2016. – Т. 2 (68), № 3. – С. 370-379.

5. Селендили Л.С., Гибадулин Р.Я. Особенности отбора и фиксации диалектного материала для «Электронного словаря крымскотатарских диалектов» // Материалы XV Всероссийской научной конференции «АКТУАЛЬНЫЕ ПРОБЛЕМЫ ДИАЛЕКТОЛОГИИ ЯЗЫКОВ НАРОДОВ РОССИИ», проводимой в рамках IV Всемирного курултая башкир и посвященной юбилею доктора филологических наук, профессора Ф.Г. Хисамитдиновой. Уфа: ФГБУН ИИЯЛ УНЦ РАН, 2015. – С. 248-251.

6. Гибадулин Я.Н., Гибадулин Р.Я., Сакаев А.Р., Саламатин И.М. Электронные словари тюркских языков. Компьютерная обработка тюркских языков. Первая международная конференция: Труды. – Астана: ЕНУ им. Л.Н. Гумилева, Астана: ЕНУ им. Л.Н. Гумилева, 2013. – С. 340.

7. Гибадулин Р.Я., Гибадулин Я.Н., Саламатин И.М. Электронные мультимедийные словари для татарского и башкирского языков. Материалы 5-й всероссийской тюркологической конференции “Урал-Алтай: через века в будущее”. Уфа. ИИЯЛ УНЦ РАН, 2012. – С. 24.