УДК 800.87(045)
DOI 10.22213/2413-1172-2018-1-104-108

Р. А. Верняева, кандидат филологических наук, ИжГТУ имени М. Т. Калашникова

ИСПОЛЬЗОВАНИЕ ЛИНГВОГЕОГРАФИЧЕСКОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ «ДИАЛЕКТ» В УЧЕБНОМ ПРОЦЕССЕ (НА ПРИМЕРЕ КУРСА «ТЕХНОЛОГИИ ОБРАБОТКИ ТЕКСТА И ЗВУЧАЩЕЙ РЕЧИ») *

Введение

Курс «Технологии обработки текста и звучащей речи» разработан для студентов направления «Фундаментальная и прикладная лингвистика» и в соответствии с ФГОС по данной дисциплине включает два блока: 1. Автоматическая обработка текста. 2. Технологии обработки звучащей речи.

В рамках практических и семинарских занятий планируется предоставить студентам на правах редактирования [1] возможность работы в базах данных доступных систем, в частности лингвогеографической информационной системе «Диалект» (далее ЛГИС «Диалект»), позволяющей обрабатывать как текстовые документы, загруженные в систему, так и анализировать звуковые файлы формата WAV [1].

ЛГИС «Диалект» – система, позволяющая хранить диалектный материал в различных формах (паспортизованные лексические данные, собранные по программе Лексического атласа русских народных говоров (далее ЛАРНГ), транскрибированные записи речи диалектоносителей, аудио- и видеозаписи разговоров с информантами), просматривать (прослушивать) записи, отмечать в текстах диалектные слова, представлять диалектную лексику на масштабируемой лингвистической карте и в виде статей электронного словаря [2].

Ранее в работе [3] были описаны цель создания, материал и возможности мультимедийной части корпуса русских говоров Удмуртии, который создается коллективом лингвистов и программистов под руководством кандидата филологических наук Е. А. Ждановой в рамках проекта «Русские говоры Удмуртии: корпус диалектных текстов второй половины 20 – начала 21 в.».

Формирование знаний об основных технологиях автоматической обработки устной речи, в частности особенностях обработки звуковых файлов диалектной речи и подготовке их к загрузке в базу данных, является одной из задач курса «Технологии обработки текста и звучащей речи». Результаты представления в мультимедийном корпусе аудиозаписей студенты-лингвисты направления «Фундаментальная и прикладная лингвистика» смогут использовать при подготовке курсовых и выпускных квалификационных работ.

Изучение возможностей ЛГИС «Диалект» осуществляется студентами данного направления также в рамках темы «Структура и контент лингвистических информационных и информационно-аналитических систем» лекционно-практического курса «Электронные издания» [4]. Данная дисциплина предполагает обучение студентов созданию и эффективному использованию электронных изданий. В связи с этим в рамках дисциплины «Электронные издания» учащиеся знакомятся со структурой и контентом ЛГИС «Диалект», при изучении курса «Технологии обработки текстов и звучащей речи» студенты подробно изучают и закрепляют на практике принципы автоматической обработки устного и письменного текстов.

Цифровое представление записей диалектной речи: обработка и подготовка к загрузке

В настоящее время существует множество программных продуктов, позволяющих представить звук в цифровом виде. Студенты могут выбрать удобную для их работы и предпочтений программу.

В рамках данной статьи описан процесс оцифровки аудиозаписей диалектной речи с по-

* Исследование выполнено при финансовой поддержке РФФИ и Правительства Удмуртской Республики в рамках научного проекта № 17-14-18006.

[1] Под правами на редактирование понимается доступ к документу с возможностью вносить изменения (добавлять, удалять информацию). Зарегистрированный в системе пользователь имеет права просмотра документа, не редактирования. Студентам при работе в системе предоставляются права редакторов.

мощью свободно распространяемого программного продукта Audacity (http://audacity-free.ru/), являющегося редактором звуковых файлов и предоставляющего возможности по обработке звука любого качества [5].

Рабочую область данного продукта условно можно разделить на три части: 1) область управления записью, где содержатся кнопки записи, прослушивания, паузы и др., различные инструменты, настройка устройств ввода; 2) диаграмма звукового файла; 3) строка состояния [6].

Для цифрового представления звуковых файлов необходимо синхронизировать воспроизведение кассеты и записи звукового файла на жесткий диск компьютера (рис. 1).

Рис. 1. Процесс оцифровки звукового файла

По окончании записи звуковых файлов на жесткий диск необходимо экспортировать звук в формат WAV (подробнее см. [7]).

Важным является полное и правильное представление параметров каждого звукового файла. Нами были определены следующие ключевые параметры паспортизации материала: 1) полные данные об информанте (полные фамилия, имя отчество; дата рождения; место рождения; образование; религиозная принадлежность; место работы); 2) место сбора материала (район, населенный пункт); 3) тип записи (рассказ, вопрос-ответ и т. д.); 4) качество записи (качество исходного файла); 5) тематика; 6) продолжительность записи (в формате «минута-секунда»); год записи (год сбора материала и фиксирования его на аудиокассете); 7) объем записи (в МБ); 8) имена собирателей (ФИО полностью).

Выделенные параметры способствуют удобному хранения оцифрованных файлов и дальнейшему быстрому их поиску в базе данных

Объем базы данных ЛГИС «Диалект»

База данных ЛГИС «Диалект» включает более 400 папок с диалектными текстами. Следовательно, в настоящее время студенты уже имеют возможность работать не только с описанием того или иного диалектного текста, но и автоматически обрабатывать фотокопию оригинальной записи, в частности осуществлять разметку текста по тематическим блокам, соответствующим разделам программы ЛАРНГ.

Объем загруженных звуковых файлов составляет 100 единиц. Оцифрованные аудиозаписи диалектной речи позволят в дальнейшем осуществить морфологическую и синтаксическую разметку текстов, проанализировать фонетические особенности речи диалектоносителей, а также уточнить данные о грамматических и фонетических чертах русских говоров Удмуртии.

Закрепление студентами лекционного материала курса «Технологии обработки текста и звучащей речи» на платформе ЛГИС «Диалект» и дальнейшее использование текстового и звукового материала корпуса может послужить базой для написания ими курсовых, выпускных квалификационных и научно-исследовательских работ.

Визуализация оцифрованного материала

Визуализация диалектного материала осуществляется с помощью форм запроса и вывода запроса (подробнее о возможностях системы см. [8]).

На домашней странице сайта лингвогеографической системы «Диалект» расположены вкладки «Тексты», «Карты», «Словарь» (рис. 2).

Для работы с оцифрованным материалом студенту необходимо нажать вкладку «Тексты». На данной странице указан перечень папок с записями, загруженными в базу данных (рис. 3).

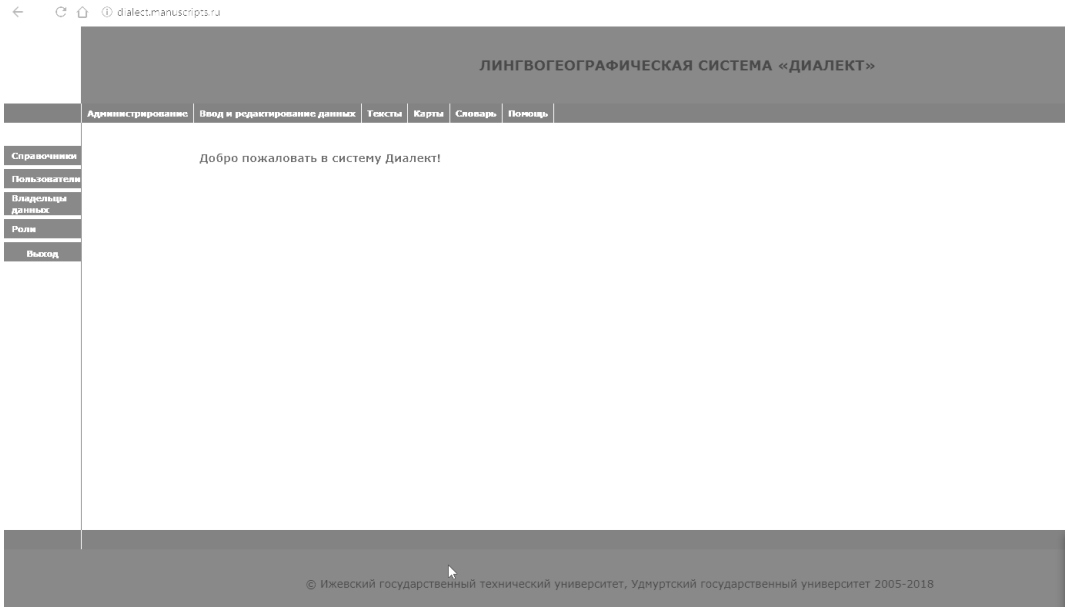
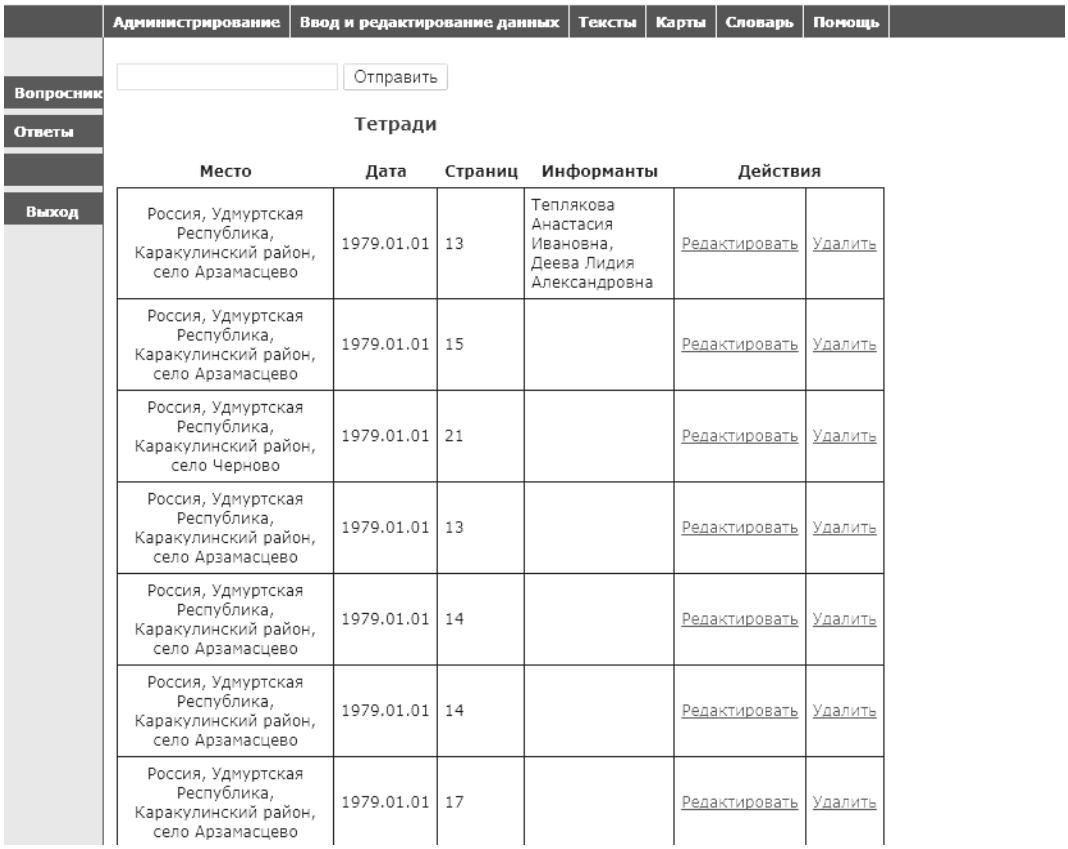


Рис. 2. Главная страница ЛГИС «Диалект»



Место	Дата	Страниц	Информанты	Действия	
Россия, Удмуртская Республика, Каракулинский район, село Арзамасцево	1979.01.01	13	Теплякова Анастасия Ивановна, Деева Лидия Александровна	Редактировать	Удалить
Россия, Удмуртская Республика, Каракулинский район, село Арзамасцево	1979.01.01	15		Редактировать	Удалить
Россия, Удмуртская Республика, Каракулинский район, село Черново	1979.01.01	21		Редактировать	Удалить
Россия, Удмуртская Республика, Каракулинский район, село Арзамасцево	1979.01.01	13		Редактировать	Удалить
Россия, Удмуртская Республика, Каракулинский район, село Арзамасцево	1979.01.01	14		Редактировать	Удалить
Россия, Удмуртская Республика, Каракулинский район, село Арзамасцево	1979.01.01	14		Редактировать	Удалить
Россия, Удмуртская Республика, Каракулинский район, село Арзамасцево	1979.01.01	17		Редактировать	Удалить

Рис. 3. Расположение тетрадей в базе данных

Для выбора текстов (письменных и аудиозаписей) необходимо нажать кнопку «Редактировать» напротив интересующей папки [2].

Далее, пройдя по ссылке, пользователь получает полный доступ ко всем оцифрованным страницам данной папки или файла с аудиозаписями. Под полным доступом мы понимаем возможность вносить изменения в метааналитические данные открытой единицы базы данных и редактирования сканированной страницы тетради (рис. 4).

Загруженные страницы имеют условное название, включающее краткое наименование населенного пункта сбора материала и порядковый номер страницы тетради (например, Арз_1 – с. Арзамасцево, страница 1).

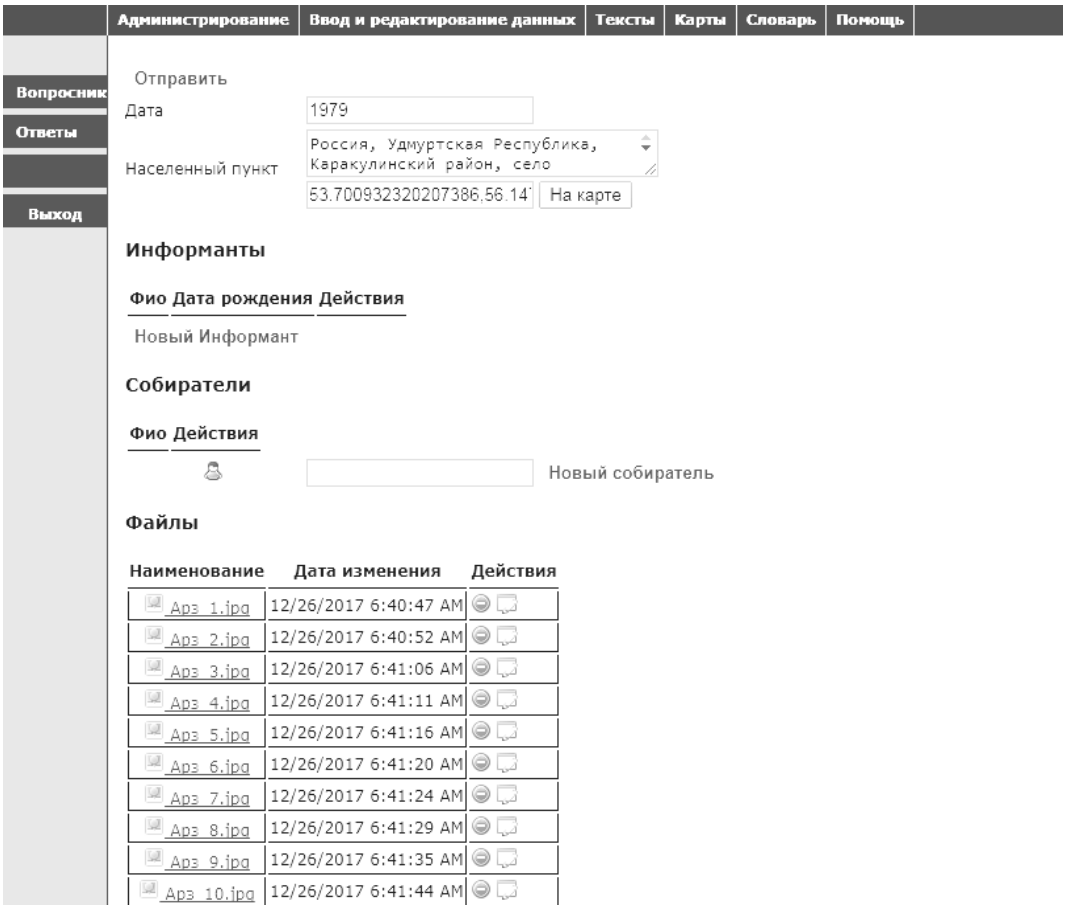


Рис. 4. Представление оцифрованного материала в базе данных

В колонке *Действия* пользователь может выбрать функцию прослушивания полного текста данной записи.

Таким образом, в настоящее время ЛГИС «Диалект» представляет собой систему, содержащую сканкопии диалектного материала и соответствующие им по содержанию оцифрованные аудиозаписи.

Выводы

Мультимедийная часть корпуса русских говоров Удмуртии в настоящее время содержит оцифрованные аудиоматерилы диалектологических практик студентов филологического факультета УдГУ в период 1994–2007 гг.

Существующие в базе данных материалы и возможности ЛГИС «Диалект» обеспечивают студентам (а также всем пользователям системы) 1) быстрый поиск необходимых для научной работы данных в удаленном режиме; 2) возможность изучения и описания фонетических, грамматических, лексических особенностей русских говоров Удмуртии; 3) закрепление теоретических знаний в области автоматической обработки письменных и устных текстов.

Библиографические ссылки

1. Лингвогеографическая информационная система «Диалект». URL: http://dialect.manuscripts.ru/ (дата обращения: 10.12.2017).

[2] Описание объекта «Папка» в БД включает следующие параметры: место и дата сбора материала, количество отсканированных страниц, персональные данные информантов.

2. Лингвогеографическая система «Диалект»: история создания, новые возможности, технологические решения, демонстрация данных / В. А. Баранов, Е. А. Жданова, Д. Б. Кожевников, А. А. Белых // Интеллектуальные системы в производстве. 2013. № 1(21). С. 171–175.

3. *Верняева Р. А.*, *Жданова Е. А.* Цифровое представление звучащей диалектной речи в лингвогеографической информационной системе «Диалект» // Гуманитарное образование и наука в техническом вузе: сборник докладов Всерос. науч.-практ. конф. с междунар. участием / отв. ред. В. А. Баранов. Ижевск : Изд-во ИжГТУ имени М. Т. Калашникова, 2017. С. 457–463.

4. *Верняева Р. А.* Электронные издания. Практические работы и рекомендации к их выполнению : метод. реком. к вып. практ. работ для студ. напр. «Фундаментальная и прикладная лингвистика» профиля «Теоретическая и прикладная лингвистика» [Электронный ресурс]. Ижевск : Изд-во ИжГТУ имени М. Т. Калашникова, 2017. 28 с. URL: https://drive.google.com/file/d/11BPWHBMY4zBRA0L0okYOik7oFYzZZx5Q/view (дата обращения: 10.12.2017).

5. Руководство по программе Audacity [Электронный ресурс]. URL: http://beginwithsoftware.com/videoguides/audacity-guide-rus.html (дата обращения: 11.12.2017).

6. *Верняева Р. А.*, *Жданова Е. А.* Указ. соч.

7. Там же.

8. Лингвогеографическая система «Диалект»: история создания, новые возможности, технологические решения, демонстрация данных…

References

1. *Lingvogeograficheskaja informacionnaja sistema "Dialect"* [Linguistic-geographical information system "Dialect"], available at http://dialect.manuscripts.ru/ (accessed December 22, 2017) (in Russ.).

Получено 29.12.2017

2. Baranov V. A., Zhdanova E. A., Kozhevnikov D. B., Belyh A. A. (2013). *Intellektual'nye sistemy v proizvodstve* [Intelligent Systems in Manufacturing], vol. 1(21), pp. 171-175 (in Russ.).

3. Vernyaeva R. A., Zhdanova E. A. (2017). *Cifrovoe predstavlenie zvuchashchei rechi v lingvogeograficheskoi informacionnoi sisteme "Dialect"* [Digital representation of sounding dialect speech in the linguistic-geographical information system "Dialect"]. Proceedings of the *Gumanitarnoe obrazovanie i nauka v tehnicheskom vuze* (ed. Baranov V. A.), pp. 457-463 (in Russ.).

4. Vernyaeva R. A. (2017). *Jelektronnye izdanija. Prakticheskie raboty i rekomendacii k ih vypolneniju: metodicheskie rekomendacii k vypolneniju prakticheskih rabot dlja studentov napravlenija "Fundamental'naja i prikladnaja lingvistika" profilja "Teoreticheskaja i prikladnaja lingvistika"* [Electronic publications. Practical work and recommendations for their implementation], available at https://drive.google.com/file/d/11BPWHBMY4zBRA0L0okYOik7oFYzZZx5Q/view (accessed December 10, 2017) (in Russ.).

5. *Rukovodstvo po programme Audacity* [Audacity guide], available at http://beginwithsoftware.com/videoguides/audacity-guide-rus.html (accessed December 25, 2017) (in Russ).

6. Vernyaeva R. A., Zhdanova E. A. (2017). *Cifrovoe predstavlenie zvuchashchei rechi v lingvogeograficheskoi informacionnoi sisteme "Dialect"* [Digital representation of sounding dialect speech in the linguistic-geographical information system "Dialect"]. Proceedings of the *Gumanitarnoe obrazovanie I nauka v tehnicheskom vuze* (ed. Baranov V. A.), pp. 457-463 (in Russ.).

7. Ibid.

8. Baranov V. A., Zhdanova E. A., Kozhevnikov D. B., Belyh A. A. (2013). *Intellektual'nye sistemy v proizvodstve* [Intelligent Systems in Manufacturing], vol. 1(21), pp. 171-175 (in Russ.).