КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

УДК 800.879

В. А. Баранов, доктор филологических наук, профессор;
Е. А. Жданова, кандидат филологических наук, доцент;
Ижевский государственный технический университет
имени М. Т. Калашникова
Д. Б. Кожевников, руководитель проектов;
ООО HeadLine, Ижевск
А. А. Белых, аспирант
Ижевский государственный технический университет
имени М. Т. Калашникова

ЛИНГВОГЕОГРАФИЧЕСКАЯ СИСТЕМА «ДИАЛЕКТ»: ИСТОРИЯ СОЗДАНИЯ, НОВЫЕ ВОЗМОЖНОСТИ, ТЕХНОЛОГИЧЕСКИЕ РЕШЕНИЯ, ДЕМОНСТРАЦИЯ ДАННЫХ

Описание основных возможностей компьютерной системы, предназначенной для хранения, редактирования, обработки и демонстрации в Интернете диалектной лексики, собранной по программе собирания сведений к Лексическому атласу русских народных говоров. Основное внимание уделено описанию решений по включению в базу данных иных форм представления данных – сканированных копий страниц с экспедиционными записями текстов и аудио- и видеоматериалов – и по их интеграции с существующими картографическими и лексикографическими функциями системы.

Ключевые слова: компьютерная лингвистика, автоматизированное картографирование и лексикографирование, диалектология, лингвистическая география

Цели и история создания системы «Диалект»

Компьютерные технологии все активнее используются в тех направлениях изучения языка и речи, где до сих преобладали ручная первичная обработка данных, их анализ традиционными методами и демонстрация результатов исследований в печатных изданиях. И применение вычислительной техники в таких областях во многом подготовлено возможностью достаточно простого и непротиворечивого представления лингвистических данных в виде формальных моделей. Сказанное в полной мере относится к работам по лингвистической географии, методика которой предполагает сбор материала по фиксированным вопросникам, паспортизацию ответов – указание места, времени фиксации и характеристик информанта, строгую систематизацию данных и нанесение их на карту с помощью условных знаков с целью выявления территорий распространения как отдельных языковых явлений, имеющих альтернативные формы или значения, так и диалектных областей, зон и групп и последующей интерпретации языкового ландшафта.

В конце 1980-х годов в Институте лингвистических исследований РАН (ИЛИ РАН) начались работы по созданию Лексического атласа русских народных говоров (ЛАРНГ, рук. – проф. И. А. Попов) Европейской части России. В 1991 году начался сбор материала по вопросникам будущего атласа [1] на территории Удмуртии (рук. – тогда доц. В. А. Баранов). К настоящему времени обследовано 198 сельских населенных пунктов из двадцати пяти районов Удмуртии, записано более 225 тыс. ответов на вопросы программы ЛАРНГ.

В конце 1990-х годов для перевода рукописного материала в электронную форму была создана локальная база данных для ввода, редактирования и хранения ответов на вопросы программы, ее вопросников и сведений об информантах, населенных пунктах и собирателях (программист – А. Н. Миронов), а в 2005–2008 годах – первая версия лингвогеографической информационной системы «Диалект» (ЛГИС «Диалект»), ориентированная на распределенный и удаленный ввод и редактирование данных, подготовку запросов и отбор материала с помощью веб-форм и на демонстрацию карт, выполненных в значковой технике, в Интернете (руководитель – проф. В. А. Баранов, программист – И. С. Соломенников) [2–5]. В 2011–2012 годах система была существенно переработана как с точки зрения модели, так и в части административных и пользовательских процедур и интерфейсов (руководитель – доц. Е. А. Жданова, программист – Д. Б. Кожевников) [6, 7].

Основные возможности

В серии статей, вышедших за восемь лет развития и эксплуатации ЛГИС «Диалект», достаточно подробно рассмотрены основные параметры и функции как первой, так второй версий системы, показаны некоторые результаты их использования (см. список библиографических ссылок). Здесь необходимо кратко остановиться на возможностях, которые предоставляются пользователям в настоящее время.

База данных. Предназначена для хранения лингвистической информации, полученной в ходе обследования русских говоров междуречья Вятки и Камы по вопросникам программы ЛАРНГ, а также текстовой информации в различных формах – сканкопий страниц тетрадей с записями диалектной речи, аудио- и видеофайлов, транскрипций. Наличие в базе данных экстралингвистической информации о времени, месте фиксации материала, об информантах и собирателях позволяет осуществлять запросы, идентифицировать данные и визуализировать их [8].

Запросная форма картографического модуля. Четырехоконный интерфейс запросной формы позволяет:

– выбрать вопрос(ы) программы, ответы на которые интересуют пользователя, и/или иные параметры имеющихся в базе данных ответов: пункт(ы), время, возраст, пол, образование информантов и др.;

– выбрать поля, значения которых необходимо визуализировать;

– указать группировку и сортировку выводимых в виде таблицы значений.

Демонстрация выборки. Визуализация выборки осуществляется:

– в виде интерактивной таблицы, которая позволяет просмотреть и выбрать для показа на карте или отвести от картографирования данные;

– в виде карты, на которую нанесены знаки, установленные пользователем в качестве соответствий отобранным для картографирования данным.

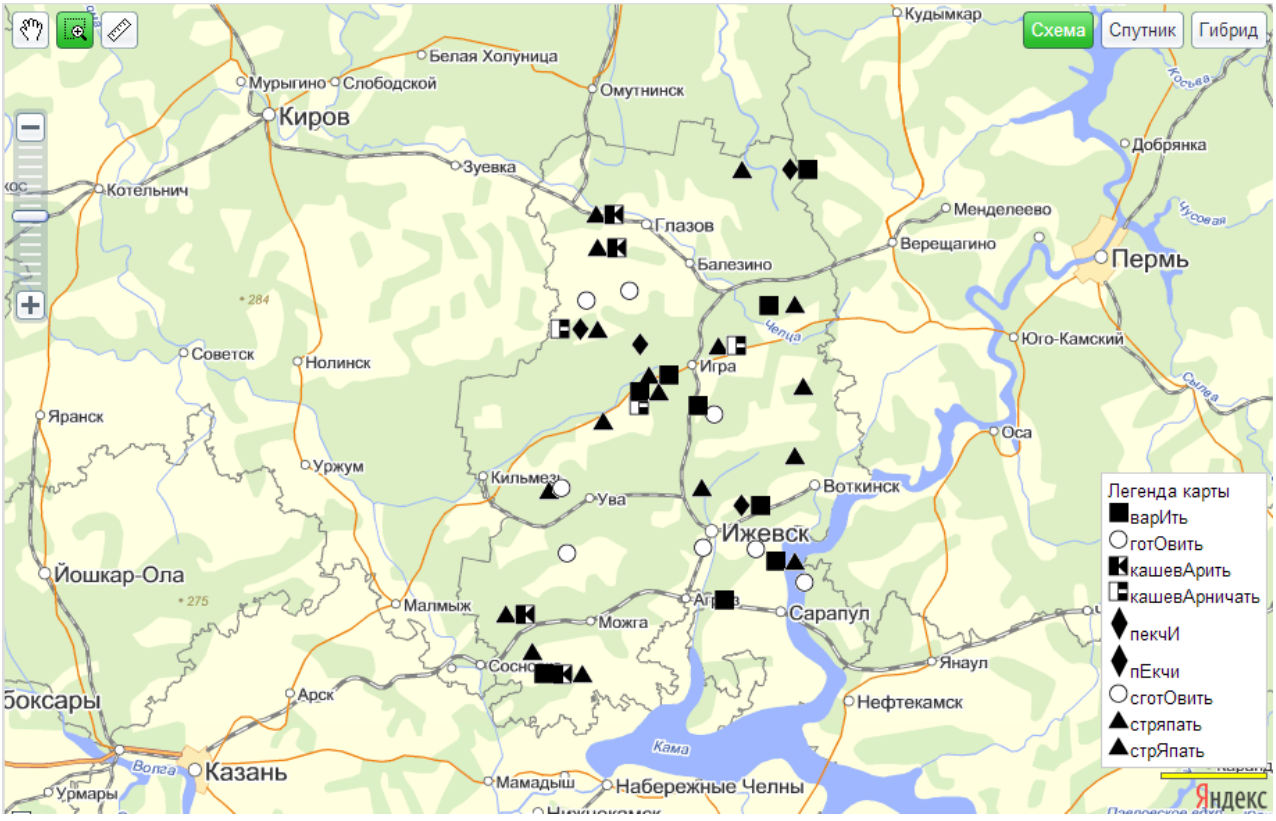
Карты. Для визуализации данных используются интегрированные с системой «Диалект» карты «Яндекс», что позволяет не только масштабировать их вывод, но и при наличии в базе данных аналогичного лексического и/или текстового материала, а также справочной информации о населенных пунктах других регионов осуществлять картографирование и лексикографическое представление диалектной лексики иных территорий [6, 9].

Запросная форма лексикографического модуля позволяет найти в базе данных слова, соответствующие маске запроса, визуализировать информацию о них в форме структурированной словарной статьи, а при необходимости перейти по гиперссылке к дополнительной информации. В настоящее время материалом для построения словарных статьей являются ответы на вопросы программы ЛАРНГ [7].

Демонстрация данных в Интернете

Картографирование

Карты, построенные при помощи ЛГИС «Диалект», позволяют визуализировать языковой материал, имеющийся в базе данных по выбранному вопросу. В правом нижнем углу представлена легенда, отражающая картографируемые слова и соответствующие им на карте условные обозначения. Данная карта позволяет увидеть разнообразие глаголов, называющих процесс приготовления пищи в русских говорах Удмуртии. Она показывает, что в некоторых населенных пунктах для обозначения этого действия используется целый ряд наименований.



Карта к вопросу 19002 «Готовить пищу (общее наименование)» раздела «Питание» программы ЛАРНГ

Повсеместно распространенным оказывается обозначение *стряпать*. Глагол *варить* как общее наименование процесса приготовления пищи преобладает в восточной части республики, а однокоренные образования *кашеварить* и *кашеварничать* – в западной. Обозначение *пекчи* в качестве общего наименования процесса приготовления пищи используется в основном в говорах северной части Удмуртии.

На подготовительном этапе от картографирования был отведен просторечный глагол *кухарить*, распространенный на всей территории Удмуртской Республики, а также не соответствующие теме карты единичные наименования *варганить, гоношить, делать*. Общеупотребительный глагол *готовить* и соответствующий ему глагол совершенного вида *сготовить* отражены на карте лишь в тех пунктах, где они являются единственными названиями обозначаемого процесса. При выборе значков для картографирования акцентологические и видовые различия не учитывались, поэтому соответствующие варианты глаголов переданы на карте одинаковыми значками.

Лексикографирование

Словарная статья системы имеет следующую структуру и состав:

лексема (ответ на вопрос программы ЛАРНГ) *(количество_фиксаций)*

текст_вопроса_программы (аналог дефиниции)

место_фиксации_слова дата_фиксации_слова контекст

место_фиксации_контекста дата_фиксации_контекста

Ср. синоним (другой ответ на тот же вопрос)

См. также омоним (вопрос, на который дан идентичный ответ).

Некоторые элементы словарной статьи имеют гиперссылки, позволяющие получить дополнительную информацию или перейти к аналогичным данным.

Приведем пример:

варИть (24)

Готовить пищу (общее название)

республика Удмуртская, Граховский район, село Грахово (2001)

Дебесский район, село Смольники (2001)

Завьяловский район, село Завьялово (1998)

Кезский район, село Степаненки (2001)

Киясовский район, село Первомайский (2000)

Сарапульский район, село Девятово (1998)

Якшур-Бодьинский район, село Порва (1998)

село Старые Зятцы (1999)

специальное

республика Удмуртская, Кезский район, село Степаненки (2001)

Киясовский район, село Первомайский (2000)

Якшур-Бодьинский район, село Варавай (2001)

село Старые Зятцы (1999)

«порА об'Эд вар'Ит'» – Россия, республика Удмуртская, Киясовский район, село Первомайский (2000)

«об'Эд нАдо итт'И вар'Ит'» – Россия, республика Удмуртская, Киясовский район, село Первомайский (2000)

Россия, республика Удмуртская, Воткинский район, село Перевозное, 1984 001.jpg

См. также: варганить варгАнить варИть гоношИть готовАть готовить готовИть готОвить дЕлать зготОвить Ись готОвить кашевАрить кашевАрничать кашовАрить кухАрить кухарничать кухАрничать куховАрить настрЯпать пекчИ пЕкчи приготовлЯть сваргАнить сварИть сготОвить сдЕлать спрАвить стряпать стрЯпать

Ср.: Варить Кипятить Приготовление пищи (общее название) Припасать

Новые возможности и функции

Не менее ценным материалом для изучения лексической системы русских говоров являются записи диалектной речи, собранные преподавателями и студентами Удмуртского государственного университета в период с 1972 по 1990 год в ходе диалектологических экспедиций в 16 районах республики: Балезинском, Вавожском, Воткинском, Граховском, Завьяловском, Каракулинском, Кезском, Кизнерском, Киясовском, Красногорском, Можгинском, Сарапульском, Селтинском, Сюмсинском, Шарканском и Якшур-Бодьинском.

Для хранения и работы с этими материалами в системе «Диалект» доработана модель базы данных, созданы интерфейсы для просмотра сканированных страниц тетрадей с записями и для подготовки транскрипций на их основе.

Новый пункт меню «Тексты» содержит таблицу, отражающую основную информацию о текстах: место и время записи, имена информантов. Гиперссылки и всплывающие подсказки позволяют просмотреть более подробные сведения об информантах (возраст, место рождения, национальность, профессия, вероисповедание, образование) и собирателях, хранящиеся в структурах метаданных базы, а также визуализировать графические файлы с текстами.

Новый модуль интегрирован с основной базой данных. Это обеспечивается как использованием тех же структур для метаописаний, что и для данных ЛАРНГ, так и благодаря функции разметки на фрагменты графического изображения страницы и вводе в базу данных транскрипции, соответствующей тексту в выделенном фрагменте. Установленные пользователем связи между графическим фрагментом, транскрипцией и вопросом ЛАРНГ позволяют включить текстовые прецеденты и контексты в виде машиночитаемых копий и в виде изображений страниц в словарные статьи и использовать их при картографировании. Так, в приведенном выше примере словарной статьи пользователь имеет возможность перейти по ссылке *Россия, республика Удмуртская, Воткинский район, село Перевозное, 1984 001.jpg* к примеру употребления слова *варить*, которому по-

священа статья, в тексте в виде сканированного изображения или его транскрипции: «*Чё нынче готовят, то и мы готовили. Чё варили? Чё придётся, то и сваришь, настряпашь к празднику только*».

Еще одной технологической инновацией системы, существенно расширяющей, в частности, ее иллюстративный потенциал, является возможность загрузки, метаописания, разметки аудио- и видеозаписей и подготовки транскрипционных записей на их основе и установления связей размеченных отрывков, в частности, с вопросами программы собирания сведений для ЛАРНГ, что позволяет включить аудио- и видеозаписи в состав словарной статьи, относящейся к заглавной лексеме, в качестве иллюстративного материала.

В целом доработка базы данных, а также административных и пользовательских интерфейсов позволила представить исходно различные по форме лингвистические ресурсы системы – ответы на вопросы программы ЛАРНГ, записи диалектной речи, аудио- и видеозаписи – как единый массив данных, лингвистические компоненты которого имеют идентичное метаописание, возможность приведения текстов и контекстов и их лингвистических единиц к единому формату транскрипции и к начальной форме слова, возможность установления связи с вопросами программы ЛАРНГ, аналогичные интерфейсы запросов, перекрестные ссылки форм вывода выборок и др.

Технологии

За пятнадцать лет работы над проектом несколько раз менялась технологическая платформа системы. Первое, локальное хранилище данных было создано с помощью СУБД Paradox под DOS. Именно в нем началась подготовка электронных копий ответов на вопросники ЛАРНГ, которые затем, после создания новой системы на базе СУБД Oracle с использованием Oracle Web Forms, были конвертированы в нее. Переход на новую платформу позволил реализовать удаленное и распределенное заполнение и редактирование базы данных и демонстрацию карт через Интернет [3].

Третья, функционирующая в настоящее время версия системы развивается с помощью программного обеспечения для создания веб-приложений ASP.NET MVC Framework и СУБД PostgreSQL 8.x, которые позволили упростить добавление новых процедур системы, снизить затраты и время на их разработку и создать мощный и гибкий пользовательский интерфейс формирования карт, вывода таблиц данных и словарных статей. Разработка системы ведется в IDE Microsoft Visual Studio 2010 на языках C# 4.0 и JavaScript, используется также pgAdmin – средства работы с СУБД Postgresql. Логически система разбита на два проекта, слой хранения/доступа к данным и пользовательский интерфейс совместно с ядром системы.

СУБД PostgreSQL 8.x и ASP.NET MVC Framework являются бесплатными платформами с открытыми исходными кодами. При разработке используются также дополнительные библиотеки, фрэймворки, программы NHibernate, Npgsql, Log4Net, jQuery, CKEditor с открытыми исходными кодами под лицензиями GPL, BSD, MSPL, MIT, Apache и их производных.

Перспективным является изменение дизайна системы с целью отказа от наследия Oracle Web Forms, а также создание мобильного приложения, обеспечивающего ввод данных, использование геолокации и запись видео- и аудиоматериалов в полевых условиях.

Таким образом, в настоящее время лингвогеографическая система «Диалект» является мощной и гибкой системой демонстрации в Интернете лингвистического материала, собранного по программе Лексического атласа русских народных говоров. Возможность хранения, демонстрации и разметки графических, аудио- и видеоматериалов и подготовленных на их основе транскрипций позволяет расширить фактографическую основу карт и продолжить создание электронного словаря русских народных говоров междуречья Камы и Вятки с привлечением текстов.

Благодарности

Работа выполнена при финансовой поддержке совета по грантам Президента Российской Федерации для поддержки молодых российских ученых-кандидатов наук, проект «Лингвистическое и программное обеспечение исследования диалектного языка методами картографирования и лексикографии», № МК–3121.2011.6.

Библиографические ссылки

1. Программа собирания сведений для Лексического атласа русских народных говоров : науч.-метод. пособие : в 2 ч. / отв. ред. И. А. Попов. – СПб. : Изд-во ИЛИ РАН, 1994.

2. *Жданова Е. А*. Применение лингвогеографической информационной системы «Диалект» в лингвистических исследованиях // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам : материалы междунар. науч. конф. (Ижевск, 13–17 июля 2006 г.) / отв. ред. В. А. Баранов. – Ижевск : Изд-во ИжГТУ, 2006. – С. 47–50.

3. *Соломенников И. С*. Лексический атлас русских говоров Удмуртии в Интернет // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам : материалы междунар. науч. конф. (Ижевск, 13–17 июля 2006 г.) / отв. ред. В. А. Баранов. – Ижевск : Изд-во ИжГТУ, 2006. – С. 161–163. URL: http://manuscripts.ru/conf/report/Solomennikov.pdf (дата обращения: 14.05.13).

4. *Жданова Е. А., Соломенников И. С*. Лингвогеографическая информационная система «Диалект» и лингвистические исследования // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам : материалы Междунар. науч. конф. (Казань, 26–30 авг. 2008 г.) / отв. ред. В. Д. Соловьев, В. А. Баранов. – Казань : Изд-во Казан. гос. ун-та, 2008. – С. 105–108.

5. *Баранов В. А., Жданова Е. А., Соломенников И. С*. Лингвогеографическая информационная система «Диа-

лект» как инструмент для представления (картографирования) диалектной лексики // Лексический атлас русских народных говоров (Материалы и исследования) 2009 / Ин-т лингвист. исслед. – СПб. : Наука, 2009. – С. 96–101.

6. *Жданова Е. А*. Лингвогеографическое изучение русских говоров Удмуртии // Лексический атлас русских народных говоров (Материалы и исследования) 2012 / Ин-т лингвист. исслед. – СПб. : Нестор-История, 2012. – С. 133–137.

7. *Жданова Е. А*. Лексикографический модуль лингвогеографической информационной системы «Диалект» // Лексический атлас русских народных говоров (Материалы и исследования) 2013 / Ин-т лингвист. исслед. – СПб., 2013 (в печати).

8. *Жданова Е. А*. Использование возможностей лингвогеографической информационной системы «Диалект» для изучения лексики русских говоров Удмуртии // Проблемы лингвистического краеведения: материалы Всерос. науч.-практ. конф., посвящ. памяти канд. филол. наук, доц. Аиды Николаевны Борисовой (г. Пермь, 12–13 окт. 2011 г.) / сост. О. В. Бражникова ; отв. ред. Ю. Г. Гладких ; Перм. гос. пед. ун-т. – Пермь, 2011. – С. 67–72.

9. *Жданова Е. А*. Анализ словообразовательных особенностей русских говоров Удмуртии при помощи лингвогеографической информационной системы «Диалект» // Информационные технологии и письменное наследие: материалы междунар. науч. конф. (Уфа, 28–31 окт. 2010 г.) / отв. ред. В. А. Баранов. – Уфа ; Ижевск : Вагант, 2010. – С. 80–85.

* * *

V. A. Baranov, Doctor of Philology, Professor, Kalashnikov Izhevsk State Technical University
E. A. Zhdanova, PhD in Philology, Associate professor, Kalashnikov Izhevsk State Technical University
D. B. Kozhevnikov, Project leader, "HeadLine" Ltd, Izhevsk
A. A. Belykh, Post-graduate, Kalashnikov Izhevsk State Technical University

Linguistic Geographical System "Dialect": History of Creation, New Opportunities, Technological Decisions, Data Demonstration

The article is devoted to description of main opportunities of a computer system intended for keeping, editing, processing and Internet demonstration of dialect vocabulary, collected by the program of compilation of information to Lexical atlas of Russian national dialects. Special attention is paid to description of decisions for including other forms of data representation into the database – scanned copies of pages with expeditionary texts notes and audio- and video-materials, and for their integration into current cartographic and lexicographical functions of the system.

Keywords: computational linguistics, linguistic automated mapping, electronic dictionaries, dialectology, areal linguistics

Получено: 26.04.13