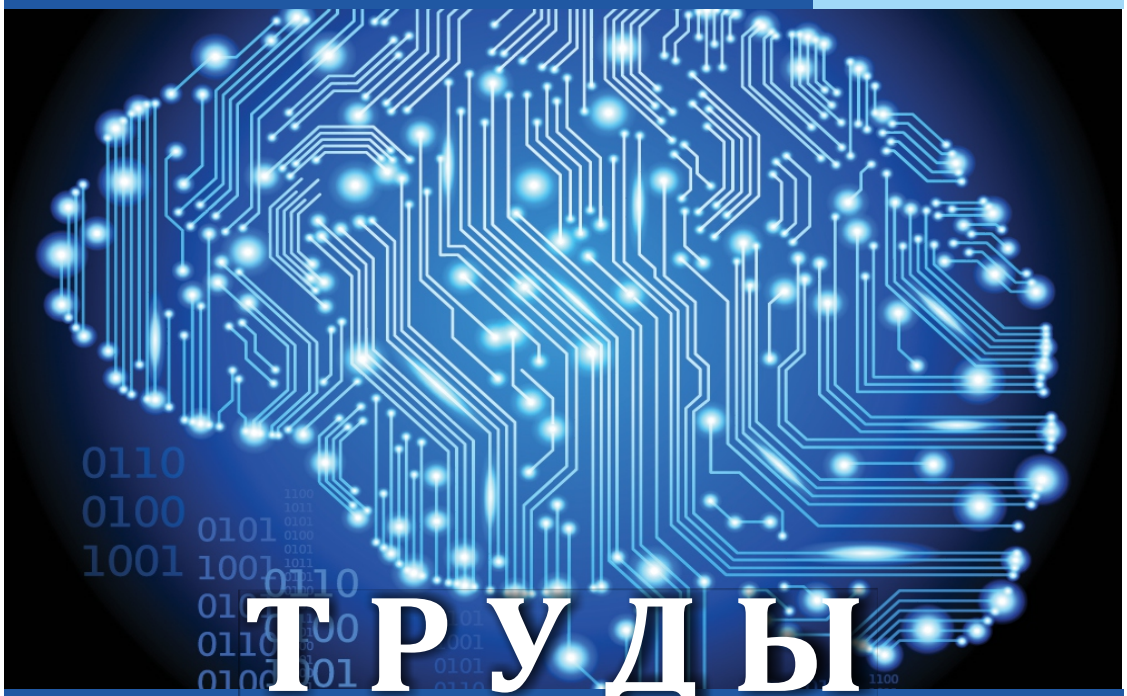
ИНТЕЛЛЕКТ
ЯЗЫК
КОМПЬЮТЕР

Выпуск
17



ТРУДЫ

МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ ПО КОМПЬЮТЕРНОЙ И КОГНИТИВНОЙ ЛИНГВИСТИКЕ

TEL–2016

Казань, 21–24 апреля 2016 г.

ИНТЕЛЛЕКТ. ЯЗЫК. КОМПЬЮТЕР

Выпуск 17

ТРУДЫ МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ ПО КОМПЬЮТЕРНОЙ И КОГНИТИВНОЙ ЛИНГВИСТИКЕ

TEL-2016

Казань, 21–24 апреля 2016

КАЗАНЬ
2016

УДК 801+681.3
ББК 81.1
Т78

Академия наук Республики Татарстан
Институт прикладной семиотики АН РТ

Казанский федеральный университет
Институт филологии и межкультурной коммуникации им. Льва Толстого
Высшая школа информационных технологий и информационных систем
Институт вычислительной математики и информационных технологий

Российский фонд фундаментальных исследований
Российская ассоциация искусственного интеллекта

Издание осуществлено при финансовой поддержке
Казанского федерального университета,
Академии наук Республики Татарстан
и Российского фонда фундаментальных исследований
(проект № 16-07-20112г).

Печатается по постановлению
Редакционно-издательского совета
Казанского федерального университета

Научные редакторы:
академик АН РТ, профессор **Д.Ш. Сулейманов**;
доцент **О.А. Невзорова**

Т78 Труды международной конференции по компьютерной и когнитивной лингвистике TEL-2016. – Казань: Изд-во Казан. ун-та, 2016. – 392 с.

ISBN 978-5-00019-650-2

Сборник содержит материалы Международной конференции по компьютерной и когнитивной лингвистике TEL-2016 (Казань, 21–24 апреля 2016).

Для научных работников, преподавателей, аспирантов и студентов, специализирующихся в области компьютерной и когнитивной лингвистики и ее приложений.

УДК 801+681.3
ББК 81.1

ISBN 978-5-00019-650-2

УДК 004.912

ЭЛЕКТРОННАЯ БАЗА ДАННЫХ АТЛАСА РУССКИХ ГОВОРОВ

А.Г.Пилюгин, Ф.И. Салимов., В.Д. Соловьев
Казанский федеральный университет, Казань
pag@kcn.ru, Farid.Salimov@kpfu.ru, maki.solovyev@mail.ru

В статье рассмотрены вопросы, связанные с созданием электронной базы данных диалектологического атласа русских говоров (ДАРЯ).

Ключевые слова: *диалекты, русский язык, ДАРЯ, базы данных*

Одним из важных аспектов исследования языковых диалектов является изучение зависимости языковых явлений от их территориального размещения. Общая среда обитания формирует определенные устойчивые связи между языками народов, населяющих данное географическое пространство, часто объясняет сходство или различие языковых явлений, которые относятся к различным языковым группам, позволяет понять характер и скорость различных изменений, происходящих в языках и диалектах. Сбор информации для фиксации соответствующих языковых явлений проводится по заранее определенной единой программе исследований в рамках заданной географической территории в течение большого отрезка времени. На базе анализа собранных материалов издаются диалектологические атласы различных языков, которые включают в себя множество карт, демонстрирующих распространение определенных языковых явлений в рамках заданных территорий. Диалектологические атласы представляют своеобразные базы данных, отражающие зависимость языковых явлений от географического положения населенных пунктов, в которых они наблюдаются.

Объединение в едином виртуальном пространстве географической и лингвистической информации позволяет ставить и решать математические задачи изучения структуры элементов этого пространства, описывать меры сходства и различия системы лингвистических признаков, характеризующих различные географические точки. Подобный подход в настоящее время приобрел достаточно широкую популярность в западных странах под названием диалектометрия. Такие исследования, прежде всего, касаются задач описания диалектного членения языков (задачи кластеризации), описания географических границ (изоглосс), разделяющих географические зоны (ареалы) распространения диалектов и говоров, выявление системы базовых признаков, наиболее значимых при

построении таких границ, описание корреляционных связей между разными языковыми явлениями. Математические исследования подобного рода особенно необходимы при решении задач большой размерности, где объемы имеющейся информации не позволяют вручную обработать большие наборы данных. Конечно, лингвисты, решая подобные задачи вручную, упрощают набор данных, выделяют и подвергают обработке главные с точки зрения исследователя компоненты. Но при этом возникают вопросы обоснованности вводимых ограничений, а также вопросы оценки точности полученных решений. Исследования по диалектометрии проводятся в ряде Европейских стран: в частности выполнены для диалектов болгарского, голландского [7-9] и др. языков. В [10] приведен обзор современного состояния диалектометрии в мире.

В течение последних двух лет в Казанском Федеральном университете совместно с лингвистами Института русского языка РАН реализуется проект создания компьютерной фактографической базы данных по говорам русского языка. Диалектологический атлас русского языка (ДАРЯ) создавался, начиная с 40-х годов XX века, усилиями большого числа исследователей и географически охватывает территорию центральных областей Европейской части России. Атлас опубликован в виде трех альбомов и трех книг сопроводительных материалов, в которые вошли сведения по фонетике, морфологии и синтаксису русского языка [1-4]. Лингвистическая информация атласа собиралась по специальной разработанной программе [5] и охватывала 294 вопроса по различным разделам языка. Обследование проводилось в специально отобранных 4209 населенных пунктах, местоположение которых образовывала относительно равномерную сетку на карте Европейской части России с шагом 15 км.

Каждая карта атласа содержит информацию по распределению значений определенных языковых явлений по населенным пунктам, в которых проводилось анкетирование. В соответствие с этим база данных разбивается на две части: картографическую, содержащую информацию о населенных пунктах, и атрибутивную, содержащую информацию по распределению значений языковых явлений по населенным пунктам. Подобная база данных для диалектов и говоров татарского языка была создана в 2012 году [11]. Для ДАРЯ электронная база данных была создана под руководством Пшеничновой Н.Н. [6]. Однако в силу имевшихся в то время различных причин не использовалась реляционная модель баз данных, для представления данных применялась оригинальная кодировка, которая на настоящий момент времени утеряна. Это обстоятельство не позволяет использовать результаты Пшеничновой и ставит задачу создания базы данных ДАРЯ заново.

Картографическая часть базы данных должна содержать информацию по населенным пунктам, в которых производился сбор информации, их географических координат, административного подчинения. База данных может также содержать дополнительную информацию по истории и этнографии, по национальному и количественному составу населения в этих населенных пунктах.

Создание картографической базы данных требует точной локализации положения каждого населенного пункта на карте. К сожалению, подобная информация в опубликованных книжных изданиях атласов отсутствует: обычно в комментариях к атласу публикуется только список наименований обследованных населенных пунктов, их административная подчиненность, без указания точных географических координат. Такое положение дел было связано с тем, что в Советское время географические координаты населенных пунктов составляли предмет государственной тайны и не могли быть опубликованы в открытой печати. Поэтому одной из задач при создании картографической базы данных была реконструкция списка населенных пунктов, приведенного в комментариях к атласу. Эта задача осложнялась длительным отрезком времени, прошедшим со времени первых полевых экспедиций. За более чем полувековой период произошли значительные изменения в составе населенных пунктов, поменялись их наименования, административная принадлежность, произошло слияние некоторых населенных пунктов, многие населенные пункты пришли в запустение и исчезли из карт соответствующих регионов. Кроме того, некоторые регионы содержат в своем составе до десяти населенных пунктов с одним и тем же наименованием, что сильно затрудняет их географическую локализацию. Особенно сложные ситуации, связанные с определением географических координат возникают при изменении административного подчинения населенных пунктов.

Для проверки и уточнения списка населенных пунктов были просмотрены существующие базы данных в Интернет. База данных http://www.bankgorodov.ru представляет открытую энциклопедию регионов, муниципальных образований и населенных пунктов России. В этой базе данных дана подробная характеристика административного устройства Российских регионов. К сожалению, многие позиции в базе данных, такие как географические координаты, численность населения, историческая и этнографическая информация для многих населенных пунктов остаются незаполненными. Другим полезным источником является электронный архив старых карт населенных пунктов Российской Федерации, расположенный по адресу http://www.etomesto.ru. База данных позволяет вести поиск географических координат населенных

пунктов, на картах, изданных несколько десятков лет тому назад. Такой ресурс особенно полезен для поиска населенных пунктов, исчезнувших с современных карт. В качестве дополнительных источников можно использовать сайт http://foto-planeta.com, в котором наряду с видовыми фотографиями населенных пунктов, содержатся сведения об их географическом положении, сайт-путеводитель http://www.esosedi.ru/, содержащий различную географическую информацию, сайт http://uistoka.ru/, содержащий информацию о исторических местах Российских регионах, различные электронные энциклопедии https://ru.wikipedia.org/, http://wikimapia.org/. К сожалению, многие из перечисленных источников формируются коллективными усилиями неквалифицированных пользователей, не имеют официального статуса, часто приводимая в них информация является неполной и противо-речивой, не отслеживается ее актуальность. Очень немногие регионы имеют официальные источники информации.

Электронные базы данных, по сравнению с их книжным аналогом обладают расширенными возможностями по представлению информации о населенных пунктах. Картографическую часть базы данных можно рассматривать как распределенную, если включить в нее ссылки на информацию, которая хранится в различных ресурсах, размещенных в сети ИНТЕРНЕТ. Подобные ссылки могут указывать на исторические, этнографические, лингвистические материалы, которые хранятся в различных базах данных и представляют интерес для исследователей диалекта и языка. К сожалению, примеров систематического описания для отдельных территорий немного. Тем не менее, очень хороший ресурс по истории и этнографии Брянской области содержится на сайте http://www.kray32.ru, некоторые интересные факты по населенным пунктам Архангельской области содержатся по адресу http://www.russia29.ru/, по населенным пунктам Владимирской области – по адресу http://vladimirskaya-rus.ru/. Понятно, что создание подобных ресурсов дело хлопотное, связанные с большими затратами, но формирование таких энциклопедических баз данных, как диалектологические атласы для отдельных языков, может в значительной мере активизировать эти процессы

Атрибутивная база данных представляет собой отображение множества языковых явлений на множество обследуемых населенных пунктов. Первичная информация собиралась в течение длительного промежутка времени (более 50 лет) большим коллективом лингвистов по программе сбора сведений для составления диалектологического атласа, принятой в 1945 году. К сожалению, исходная картотека данных во время пожара в ИРЯ РАН была утеряна. В этих условиях при реконструкции атрибутивной части базы данных приходится ориентироваться на вторичные данные, которые опубликованы в картах атласа. Отметим, что

каждая карта ДАРЯ представляет результат определенной обработки первичных данных, в них значения лингвистических признаков приписаны не населенным пунктам, а целым областям, включающим по нескольку населенных пунктов. Понятно, что подобная факторизация данных может сказаться на точности обработки результатов анализа.

Для считывания информации из карт атласа была разработана специальная процедура, которая сканирует, а далее анализирует информацию результатов сканирования. При этом приходится решать задачи, связанные с идентификацией географических координат границ приведенных на картах областей, с определением типа использованной проекции на картах атласа.

К настоящему времени завершено создание атрибутивной базы данных по первому выпуску ДАРЯ (Фонетика). Просмотрена и отредактирована база данных из 2500 населенных пунктов (всего их – 4209).

Благодарности: Работа выполнена при финансовой поддержке РГНФ (проект № 15-04-12008)

Литература

1. Диалектологический атлас русского языка. Центр Европейской части СССР. Выпуск I: Фонетика / Под ред. Р. И. Аванесова и С. В. Бромлей. — М.: Наука, 1986.

2. Диалектологический атлас русского языка. Центр Европейской части СССР. Выпуск II: Морфология / Под ред. С. В. Бромлей. — М.: Наука, 1989.

3. Диалектологический атлас русского языка. Центр Европейской части России. Выпуск III: Карты (часть 1). Лексика. — М.: Наука, 1997.

4. Диалектологический атлас русского языка. Центр Европейской части России. Выпуск III: Карты (часть 2). Синтаксис. Лексика. — М.: Наука, 2005.

5. Программа собирания сведений для составления диалектологического атласа русского языка. М.-Л., 1947.

6. Пшеничнова Н.Н. Типология русских говоров. М.: Наука. 1996. 208 с.

7. John Nerbonne and Peter Kleiweg. Lexical Distance in LAMSAS. In: John Nerbonne and William Kretzschmar (eds.) Computational Methods in Dialectometry. Special issue of Computers and the Humanities, 37(3), 2003, 339-357.

8. *Nerbonne, J. y Kretzschmar, W., 2003 , "Introducing Computational Techniques in Dialectometry" ,* Computers and the Humanities *, vol. 37 , pp. 245-255 .*

9. Houtzagers, Peter , Nerbonne, John and Prokić, Jelena (2010) 'Quantitative and Traditional Classifications of Bulgarian Dialects Compared', Scando-Slavica, 56: 2, 163 — 188

10. John Nerbonne and William Kretzschmar, Jr.Dialectometry++. LLC: Journal of Digital Scholarship in the Humanities 28(1), 2013, pp.2-12. doi:10.1093/llc/fqs062

11. Салимов Ф.И., Рамазанова Д.Б., Пилюгин А.Г., Салимов Р.Ф. Электронная версия атласа татарских народных говоров // Вестник ТГПУ 4(26), 2011, Казань, изд-во КФУ, с.205-210.