КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

УДК 800.879

Е. А. Жданова, кандидат филологических наук, доцент
А. А. Белых, аспирант
Ижевский государственный технический университет имени М. Т. Калашникова

ГЕОГРАФИЧЕСКИЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ В ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

В статье представлены основные возможности лингвогеографической системы «Диалект» в сопоставлении с характеристиками иных существующих систем картографирования лингвистических явлений и представления в Интернете лингвистических карт. Современные лингвогеографические системы вписаны в контекст развития мировой лингвистической географии.

Ключевые слова: компьютерная лингвистика, географические информационные системы, лингвистическая география, диалектологический атлас.

Представление о связи развития языка с особенностями его территориального размещения появилось в европейской лингвистике во второй половине XIX в. На основе этого представления возникло новое научное направление – лингвистическая география, – которое стало быстро набирать популярность.

В начале XX в. появляются первые диалектологические атласы – собрания карт, отражающих распространение того или иного языкового явления на определенной территории. Уже первые лингвистические атласы показали, насколько неоднородны языки, лингвогеография дала языковедам новый материал для исследований и поставила новые вопросы, поэтому работа по составлению лингвистических, в том числе диалектологических, карт активно велась во всем мире на протяжении XX в.

Открытия, сделанные в результате многолетнего картографирования языковых явлений, подтверждают, что географический – один из основных экстралингвистических факторов, влияющих на развитие и формирование языков и диалектов: естественные географические преграды часто становятся границами языковых образований, языки сопредельных территорий неизбежно оказывают влияние друг на друга. Данные, полученные лингвогеографами, используются не только языковедами, но и историками, этнографами, социологами для изучения истории заселения той или иной территории, формирования и современного состояния социальной общности, проживающей на ней.

Изначально предметом картографирования становились в основном фонетические и грамматические явления, т. к. они считались более достоверным и показательным материалом для исследования, чем лексические особенности: словарный состав языка более изменчив и в большей степени подвержен проникновению заимствуемых элементов, чем остальные. Однако впоследствии в лингвистической географии были выработаны методы сбора и анализа лексических явлений, поэтому современная лингвистическая география ориентирована в первую очередь на картографирование лексических (а также словообразовательных и семантических) диалектных различий.

В течение прошлого века было проведено лингвогеографическое исследование большинства европейских языков и частичное изучение языков других континентов. Были составлены лингвистические атласы, отражающие распространение различных языковых образований в отдельных странах или регионах. Во второй половине XX в. появляются проекты наднациональных атласов, представляющих распространение языковых явлений на территории бытования нескольких языков.

Вместе с содержательным наполнением атласов эволюционирует и техника создания лингвистических карт. Меняются и методы представления лингвистического материала на карте, и собственно технологии, применяемые при их составлении. В конце XX в. в лингвистической географии, как и во многих других областях науки, начинают применяться информационные системы. Современные компьютерные технологии, во-первых, позволяют осуществить интернет-публикацию созданных ранее атласов, что дает широкому кругу пользователей доступ к лингвистическим данным, во-вторых, они могут быть использованы в самом процессе картографирования, что облегчает работу лингвогеографов.

В качестве примера использования информационных технологий для публикации существующих атласов можно назвать проект «Электронный диалектологический атлас русского языка» [1], где на сегодняшний день представлены электронные версии ряда морфологических карт Диалектологического атласа русского языка, работа над которым велась научными коллективами нескольких вузов России в середине XX в. Часть составленных карт была опубликована в печатном виде. Как пишут разработчики проекта «Электронный диалектологический атлас русского языка», помимо интернет-публикации карт предполагается также обеспечить эту электрон-

ную систему функцией автоматизированного построения ареалов. Программа написана в среде разработки Borland Delphi 6, для хранения исходных данных диалектологических исследований используется база данных под управлением СУБД MS Access. Картографические материалы хранятся в формате *.dwg (векторный формат системы Autodesk AutoCad), для работы с данным форматом используется триальная версия компонента CAD Import VCL (фирмы Soft Gold Ltd) [2].

Другим значительным, на наш взгляд, лингвогеографическим проектом является Электронный атлас татарских народных говоров [3]. Электронная версия построена на основе материалов опубликованных в 1989 г. двух томов атласа татарских народных говоров Среднего Поволжья, Приуралья и Сибири и на базе неопубликованных материалов к третьему тому. Электронный атлас состоит из 215 карт языковых явлений, вспомогательных карт и сводных карт изоглосс. По желанию пользователя открывается вопрос, на который отвечает карта, легенда, комментарии, информация о настройках, данные о населенных пунктах. В зависимости от цели обращения к электронному ресурсу пользователь сам может выбирать вид представления материала. Атлас снабжен и средствами навигации по картам. Создание электронной версии атласа татарских народных говоров позволило выделить ряд новых переходных говоров, обнаруженных на территориях, которые являлись границами в печатных томах атласа. В отличие от печатного, электронный атлас содержит обновленный библиографический список, включающий и интернет-источники [4].

В последние десятилетия информационные технологии повсеместно используются и для составления лингвистических атласов. В современной России ведется активная работа по составлению лексических атласов русских народных говоров. Словарный состав русского диалектного языка оказался наименее обследованным в лингвогеографическом аспекте, как отмечают составители проекта лексического атласа русских народных говоров (ЛАРНГ), «из общего количества известных по областным картотекам и словарям диалектных слов (около 250 тыс.) изоглоссы определены едва ли для 1 % слов» [5, с. 5]. Еще на начальных этапах работы над ЛАРНГ (1994 г.) составители ориентировали собирателей диалектного материала на обработку данных для ЭВМ [6]. Однако эта задача не была решена в полном объеме. В 2000-х годах для составления карт ЛАРНГ начали использовать геоинформационную систему MapInfo, которая дает возможность отображать на масштабируемой карте выбранной территории в соответствии с заданными координатами лингвистическую информацию в виде установленных составителем значков, позволяет хранить картографированные данные, редактировать карты и выводить их на печать в нужном пользователю формате. К сожалению, база данных ЛАРНГ на сегодняшний день не интегрирована с MapInfo, поэтому ввод данных для картографирования осуществляется вручную, что занимает большое количество времени.

Сейчас база данных ЛАРНГ, изначально созданная для хранения и обработки ответов на вопросы программы атласа и представляющая собой электронную картотеку, содержащую более 3 млн карточек, совершенствуется в соответствии с современными запросами лингвистов-диалектологов: для облегчения процесса составления карт атласа материалы нужно преобразовать в формат, отвечающий требованиям программы MapInfo, для удобства обработки данных картотеку планируется снабдить функцией полуавтоматического ввода данных и удаленного редактирования, для осуществления поиска материала при помощи СУБД Oracle база данных будет обеспечена возможностью отбора по задаваемым пользователем параметрам (регион, тема, вопрос и т. д.) [7].

Опыт ЛАРНГ дал многим научным коллективам толчок к составлению региональных лексических атласов при помощи современных компьютерных технологий, в частности той же MapInfo. Таким образом идет работа, например, над атласом говоров Ярославской области [8].

В настоящее время MapInfo используется также для составления Лингвистического атласа прибалтийско-финских языков, Лингвистического атласа Европы, на основе материала для которого в Удмуртском государственном университете создается Диалектологический атлас удмуртского языка. Все эти атласы ориентированы на печатное издание, хотя их составители не исключают перспективу интернет-публикации.

Как пишут составители Диалектологического атласа удмуртского языка, в рамках исследования были созданы карты и геоинформационные базы данных по произношению удмуртами картографируемых слов. С целью географической локализации лексем и их форм было проведено геокодирование. Для этого была подготовлена геоинформационная база поселений – опорных пунктов сбора информации. Используемые географические информационные технологии позволяют построить различные виды диалектологических карт, совмещать значковый способ с заливкой ареалов. Посредством ряда математико-статистических процедур полученные материалы были сопоставлены между собой и найдены географические закономерности распространения лексем [9].

В то же время отметим, что работа с системой MapInfo зачастую оказывается сложной для неподготовленных пользователей, какими являются диалектологи старшего поколения, ведущие лингвогеографическую деятельность в регионах, это значительно снижает ее популярность среди российских лингвистов. К недостаткам использования MapInfo можно добавить отсутствие возможности удаленного доступа к файлам базы данных и самой системе картографирования. Составители ЛАРНГ при работе в этой системе столкнулись также с невозможностью отображения на карте нескольких значений

для одной точки. Однако при картографировании лексического материала, когда велика вероятность лексической и словообразовательной синонимии, т. е. наличия в одном и том же говоре нескольких обозначений одной реалии, составитель карты оказывается в затруднительном положении: система ограничивает возможности визуализации материала, принуждая отказаться от выведения на карту части информации.

Ряд научных коллективов, которые начали работу над лингвистическими атласами еще в XX в., в настоящее время выбрали путь, совмещающий оба представленных выше варианта электронной работы с лингвогеографическими изданиями. Так, коллектив составителей Атласа говоров Самарского края начал работу над проектом, целью которого является создание электронных версий составленных ранее карт изданного атласа, а также составление при помощи компьютерных технологий и интернет-публикация новых лексических карт [10].

Создатели Общеславянского лингвистического атласа (ОЛА), также в последние годы использующие для работы над картами систему MapInfo, приняли решение опубликовать в Интернете составленные ранее карты, материалы к ним и посвященные им публикации. Для этого в Институте русского языка РАН был создан сайт Общеславянского лингвистического атласа [11], который предоставляет возможность ознакомиться с картами опубликованных выпусков ОЛА, диалектными материалами, собранными на территории России, дает информацию об истории и ходе работы над атласом, а также сопутствующие сведения: программу сбора материала и научные работы составителей, посвященные картографированию лингвистических явлений.

Некоторые научные коллективы разрабатывают собственные лингвогеографические системы, призванные служить для картографирования говоров отдельного региона. Таким путем пошли, например, диалектологи Волгоградского государственного педагогического университета, разработавшие проект «Лексический атлас Волгоградской области» [12]. В его основе лежит материал, собранный на территории указанного региона для Лексического атласа русских народных говоров. Проект предполагает автоматическое представление вводимых диалектных данных на встроенной в систему карте Волгоградской области, возможно создание карты по каждому вопросу программы ЛАРНГ с легендой и комментариями. В качестве достоинств проекта разработчики называют простоту и открытость для пользователей и вновь поступающей информации [13]. К сожалению, в этом проекте сохраняется порайонная сетка обследований, установленная для ЛАРНГ, что мешает сделать лингвистическую карту региона более детальной, указать более четкие ареалы диалектных явлений, отсутствуют подробные метаданные о каждом фиксируемом слове (пункт и дата записи, сведения о собирателе и информанте и т. д.).

Собственная лингвогеографическая система для хранения и картографического представления диалектного материала была создана и в Удмуртском государственном университете еще в 2005 г. Изначально с помощью СУБД Paradox была создана база данных, содержащая ответы на вопросы программы ЛАРНГ, далее в результате перехода на СУБД Oracle система была функционально обогащена: появилась возможность удаленного доступа к данным и создания карт в Интернете [14].

Современная версия лингвогеографической информационной системы «Диалект» [15], созданная в 2011–2012 годах в Ижевском государственном техническом университете, была задумана как многофункциональное средство хранения, анализа и представления в Интернет диалектных данных [16].

Усовершенствованная ЛГИС «Диалект» дает возможность пользователю в соответствии с задачами исследования составить запрос, содержащий параметры, которым будет соответствовать представляемый на карте лингвистический материал (состав лексем, временной интервал фиксации данных, характеристики информантов, вид значков и т. д.), осуществляет сортировку и группировку согласно заданным критериям, позволяет выбирать значки для представления диалектных данных на карте. Созданные выборки данных и карты сохраняются системой, при необходимости пользователь может вернуться к любому этапу составления карты и внести изменения.

Так, например, в отличие от лингвогеографических систем, основанных на программе ЛАРНГ, ЛГИС «Диалект» позволяют сделать выборку и создать карту по нескольким вопросам программы, количество которых задается непосредственно пользователем. В качестве примера рассмотрим карту, на которой отражены словообразовательные особенности лексем к вопросам № 0112605 «Одна ягода клюквы» и № 0112606 «Одна ягода малины» (рис. 1). На карте отчетливо видны ареалы распространения диалектных наименований. Материал карты свидетельствует о том, что на территории Удмуртской Республики для обозначения клюквы используются следующие лексемы: *клю́ква, клю́ковинка, клю́ковка*. Что касается наименований, обозначающих ягоду малины, то здесь наряду с общерусским словом *мали́на* используются лексемы *малининка* и *мали́нка*. Лингвогеографический анализ показал, что в основном для наименования ягод в русских говорах Удмуртии используются общерусские названия *клю́ква* и *мали́на*. Однако в восточной части республики зафиксированы дериваты, образованные с помощью суффикса -к- (*клю́ковка, мали́нка*).

Лингвогеографический модуль ЛГИС «Диалект» на сегодняшний день находит наиболее широкое применение в научных исследованиях, с его помощью авторами данной статьи и студентами специальности «Теоретическая и прикладная лингвистика» ИжГТУ имени М. Т. Калашникова составлено несколько сотен карт по темам «Природа» (разделы «Животный мир» и «Растительный мир»), «Трудовая

деятельность», «Питание», «Материальная культура», отражающих лексические и словообразовательные особенности диалектов междуречья Камы и Вятки, что является основательным заделом для составления атласа русских говоров Удмуртии. Данные лингвистических карт, составленных с помощью ЛГИС «Диалект», позволяют сделать вывод о существовании на территории республики южной, западной и восточной диалектных зон, противопоставленных на большинстве полученных карт, и о наиболее значительном своеобразии говоров восточных районов республики. В качестве примера можно представить карту к вопросу № 0119700 «Лось» (рис. 2). На данной карте представлены лексико-словообразовательные различия в наименовании лося на территории Удмуртии. Лингвогеографический анализ карты показал, что в основном для наименования лося в Удмуртии используется общерусская лексема *лось*, однако в восточной части республики распространены лексемы *соха́тый, соха́ч*, а на юге зафиксирована лексема *труба́ч*. Проанализировав данные карты, можно выделить изоглоссу, идущую с северо-востока на юго-запад.

Важной чертой, отличающей ЛГИС «Диалект» от других современных лингвогеографических ресурсов, является неограниченность территории, которая может быть представлена на карте. Если большинство электронных проектов рассчитаны на представление территории, условно отграниченной составителями, то в основе лингвогеографического модуля ЛГИС «Диалект» лежит система карт «Яндекс», которая позволяет картографировать материал любого языка или говора.

Отметим также, что в отличие от большинства систем, появившихся под влиянием Лексического атласа русских народных говоров, ЛГИС «Диалект» рассчитана на соотнесение выбранного составителем значка, обозначающего диалектное явление, с конкретной точкой на карте – населенным пунктом, а не условной отметкой района, что обеспечивает более подробное представление диалектного ландшафта на географической карте.

Помимо этого, ЛГИС «Диалект» отличается от существующих сегодня аналогичных систем богатством и разнообразием форм содержащегося в базе данных материала: это не только лексика, полученная в результате опроса по программе ЛАРНГ, но и транскрипции и аудиозаписи диалектной речи, сделанные в ходе диалектологических экспедиций студентами и сотрудниками УдГУ в 1970–2000-е гг. Столь же разнообразный диалектный материал представлен сегодня только в Трансдунайском электронном корпусе [17], который является собранием карт, отображающих ареалы распространения того или иного говора болгарского языка, в сопровождении описания фонетических и лексических особенностей каждого из них. В разделе «Тексты» представлены диалектные материалы, собранные на территории Трансдуная в 1961–1975 гг. К некоторым текстам прикреплены аудиофайлы с записью диалектной речи. Кроме того, открыв страницу, посвященную какому-либо говору, можно увидеть карту его распространения, данные о собирателях и информантах и другие экстралингвистические сведения.

ЛГИС «Диалект», в отличие от всех представленных ранее электронных ресурсов, снабжена также лексикографическим модулем, который является инструментом для создания электронного диалектного словаря, базирующегося на разнообразном и значительном по объему диалектном материале, с системой поиска и перекрестных ссылок. Все лексемы, входящие в электронный словарь, паспортизированы: для каждой указывается дата, место и количество фиксаций, омонимы и синонимы, если таковые имеются, примеры употребления, имеющиеся в базе данных.

Лексикографический модуль предоставляет пользователям возможность поиска слов по любой части (корневой, аффиксальной морфеме), благодаря чему могут быть проанализированы отдельные пласты лексики, словообразовательные типы и словообразовательные гнезда. Таким образом, с помощью ЛГИС «Диалект» был проведен анализ существительных, выражающих абстрактное значение, и существительных со значением лица. Исследование словообразовательных особенностей существительных на *-ость* и *-ота*, проведенное с помощью ЛГИС «Диалект», позволило выявить черты, присущие данной категории лексики в русских говорах Удмуртии. Так, в результате анализа выяснилось, что синонимия существительных на *-ость* и *-ота* характеризуется высокой степенью семантического сходства синонимов. Субстантивы, образованные с помощью суффикса *-ость* (*ли́хость* 'злоба, злость', *недви́жимость* 'паралич', *непра́вность* 'бедность', *ди́кость* 'глупость', *многоде́тность* 'большая семья' и др.), получили в русских говорах Удмуртии более широкое распространение, что, возможно, объясняется влиянием литературного языка. Однако следует отметить, что в отличие от русского литературного языка в говорах также имеется довольно большое количество имен существительных на *-ота* (*мокро́та* 'сорная трава в огороде', *басота* 'женские украшения (о.н.)', *красно́та* 'красная смородина (куст и ягода)', *мошкота́* 'насекомые (о.н.)' и др.).

Что касается анализа существительных со значением лица, то были исследованы субстантивы, образованные при помощи суффиксов *-тель* и *-щик*. Материал говоров показал, что наиболее активной является словообразовательная модель с суффиксом -щик *(засе́вщик* 'сеяльщик', *карау́льщик* – 'сторож, охраняющий огород', *ко́вщик* 'кузнец', *огра́бщик* 'грабитель, грабительница' и др.). В русских говорах Удмуртии встречается большое количество дериватов, образованных таким способом. Также данные ЛГИС «Диалект» позволили сделать вывод, что наиболее продуктивным является образование существительных от глагольных основ на -и, а также от основ на -а/-я.

При доработке ЛГИС «Диалект» в 2011–2012 гг. усложнилась схема использования системы: включение нескольких вариантов составления карты сдела-

ло путь к ней более длительным и менее интуитивным. Но эта проблема, которая возникает при работе с большинством современных информационных систем, решается путем разработки встроенных инструкций и всплывающих подсказок.

В качестве замечаний, адресованных разработчикам, называют цветной фон карт «Яндекс», отражающий особенности рельефа и инфраструктуры картографируемой местности. Однако можно отметить, что, во-первых, для карты небольшого региона такой фон не является помехой – достаточно крупные значки не сливаются с ним, во-вторых, использование карт «Яндекс» делает фон настраиваемым по желанию пользователя, а в-третьих, для ряда карт, тема которых связана, например, с ландшафтом или транспортом, является важным представление об особенностях рельефа и инфраструктуры той местности, где зафиксированы или, наоборот, не отмечены обозначения той или иной реалии.

У некоторых специалистов вызывает вопросы и «зависимость» ЛГИС «Диалект» от сервиса «Яндекс», однако, с одной стороны, эта система является стабильной, т. к. используется повсеместно, с другой стороны, она достаточно мобильна, чтобы отображать изменяющиеся параметры и вносить необходимые пользователю дополнения (например, отмечать исчезнувшие населенные пункты, где был записан диалектный материал).

Таким образом, ЛГИС «Диалект» является одним из наиболее многофункциональных современных средств работы с диалектным материалом. Она может быть использована не только для лингвистических исследований, но и в работах историков, этнографов, археологов, т. к. диалектология тесно связана с этими науками.

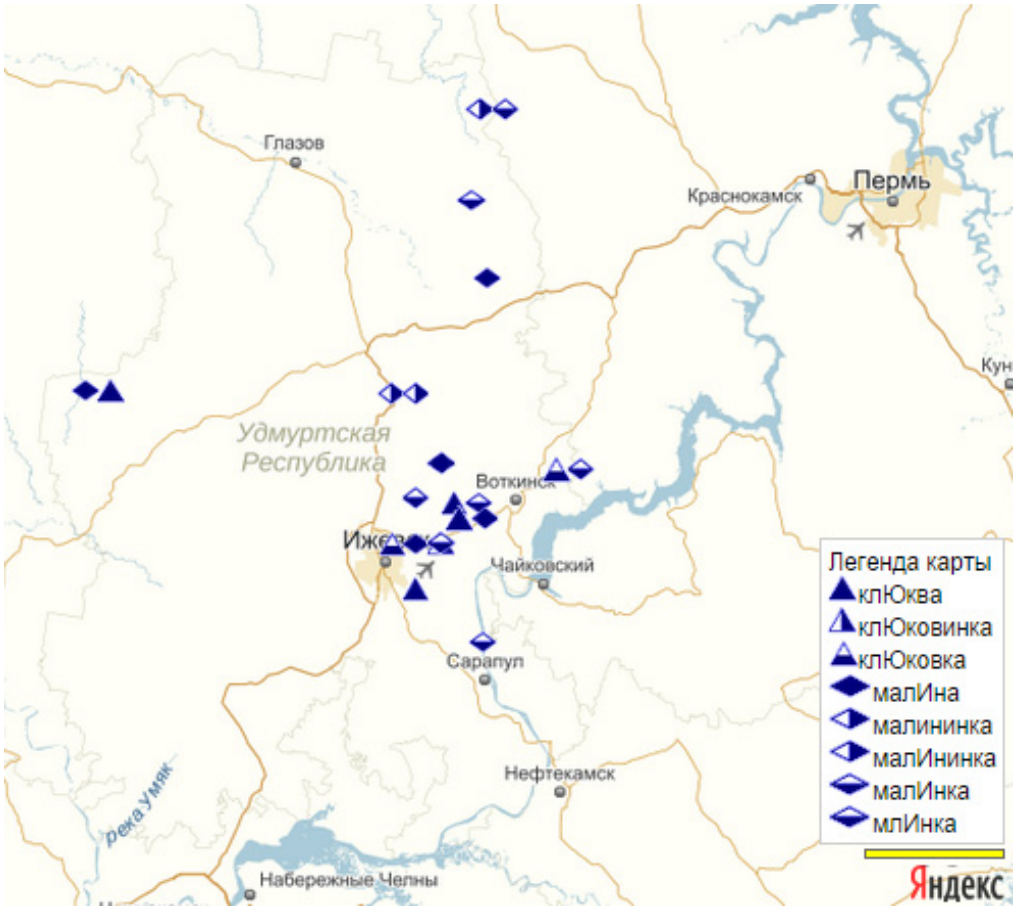

Рис. 1

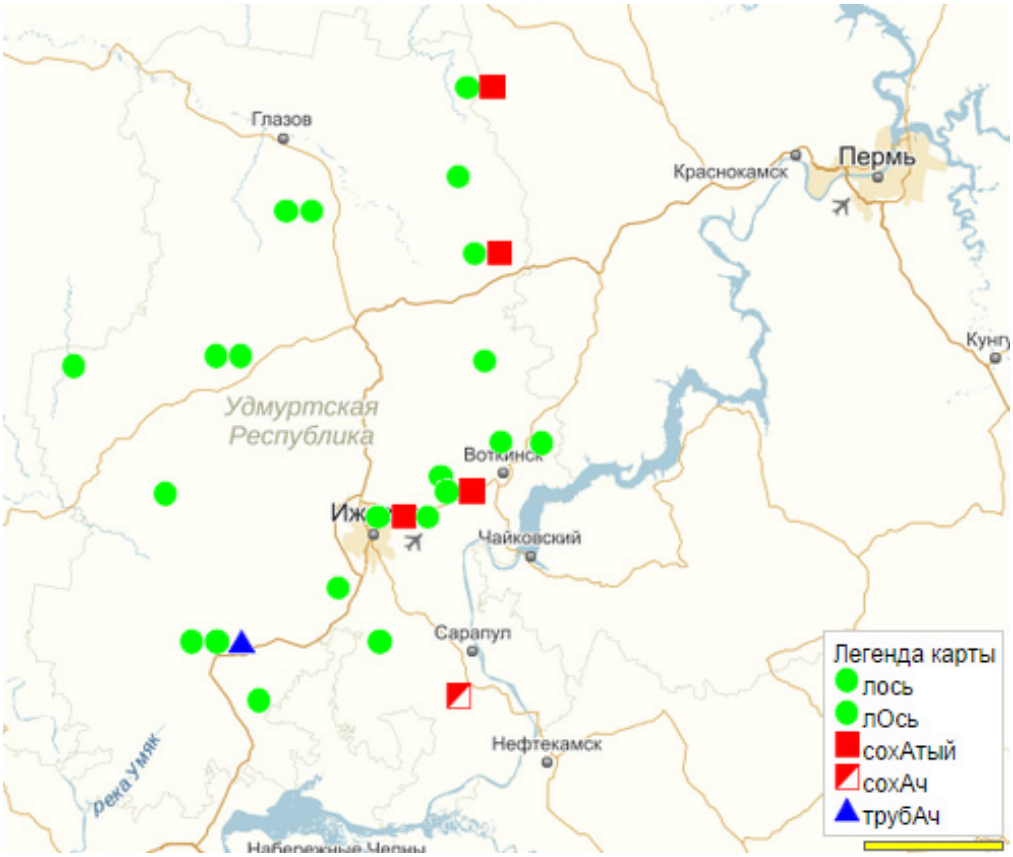

Рис. 2

Библиографические ссылки

1. Электронный диалектологический атлас русского языка [Электронный ресурс]. – URL: http://it-claim.ru/ Persons/ Kononkov/WWWroot/klaster.html.

2. Там же.

3. Электронный атлас татарских народных говоров [Электронный ресурс]. – URL: http://www.atlas.antat.ru.

4. Там же.

5. Лексический атлас русских народных говоров. Проект. / отв. ред. И. А. Попов. – СПб. : изд. ИЛИ РАН, 1994. – 112 с.

6. Там же. С. 98–100.

7. *Глебова О. В., Чихачев К. Б.* Разработка электронной картотеки для ЛАРНГ // Лексический атлас русских народных говоров (Материалы и исследования) 2010, Ин-т лингв. исслед. – СПб. : Наука, 2010. – С. 72–74.

8. *Ховрина Т. К.* Лингвогеографическое изучение лексики Ярославских говоров // Ярославский педагогический вестник. – 2009 . – № 3 (60). – С. 184–187. [Электронный ресурс]. – URL: http://vestnik.yspu.org/releases/2009_3g/ 43.pdf.

9. *Рублева Е. А., Саранча М. А.* Геоинформационные технологии в картографировании диалектов удмуртского языка [Электронный ресурс]. – URL: http://www.geogr.msu.ru/cafedra/karta/anniversary/docs/rubleva.pdf.

10. *Баженова Т. Е.* Лексика самарских говоров в ареально-типологическом аспекте [Электронный ресурс]. – URL: http://www.ssc.smr.ru/media/journals/izvestia/2014/2014_2_145_150.pdf.

11. URL: http://www.slavatlas.org.

12. URL: http://dialekt.vspu.ru.

13. *Кузнецова Е. В.* Интернет-проект «Лексический атлас Волгоградской области». Современный способ обработки диалектного материала // Лексический атлас русских народных говоров (Материалы и исследования) 2011, Ин-т лингв. исслед. – СПб. : Наука, 2011. – С. 69–80.

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

15. URL: .http://lgis2.office.hlcompany.ru.

16. *Баранов В. А., Жданова Е. А., Белых А. А.* Указ. соч.

17. URL: http://corpusbdr.info

* * *

E. A. Zhdanova, PhD in Philology, Associate Professor, Kalashnikov Izhevsk State Technical University
A. A. Belykh, Post-graduate, Kalashnikov Izhevsk State Technical University

Geographic information systems in linguistic research

The article presents the main features of linguogeographical system "Dialect" in comparison with other systems of mapping of linguistic data and presentation of linguistic maps in Internet. Modern linguogeographical systems are integrated into the context of the development of the world linguistic geography.

Keywords: computational linguistics, geographic information systems, linguistic geography, dialect atlas.

Получено: 02.10.14