ЭЛЕКТРОННЫЙ АТЛАС ТАТАРСКИХ НАРОДНЫХ ГОВОРОВ КАК ИНСТРУМЕНТ ИССЛЕДОВАНИЯ[1]

Ф.И. Салимов, Казанский Федеральный (Приволжский) университет,
Институт прикладной семиотики АН РТ, Farid.Salimov@ksu.ru
А.Г. Пилюгин, Казанский Федеральный (Приволжский) университет, pag@kzn.ru,
С.А.Ершов, ГКУЗ «Республиканское бюро судебно-медицинской экспертизы МЗ РТ»,
esa943@mail.ru

Рассматриваются типы запросов к базе данных электронного атласа татарских народных говоров. Показано, что представленная в атласе информация может быть интерпретирована как подмножество точек трехмерного пространства, осями которого являются соответственно набор населенных пунктов, в которых проводились исследования говоров, набор значений языковых явлений, и набор карт атласа, каждая из которых отражает некоторое языковое явление. Любой запрос реализует выбор подмножества точек этого пространства. Кроме того по каждой из осей возможны группировки.

С развитием информационных технологий все большее число результатов научных исследований переводится в электронный формат и становится доступным большому числу исследователей благодаря сети ИНТЕРНЕТ. Одним из существенных преимуществ электронных средств описания изучаемых объектов перед его традиционными аналогами представления является возможность использования средств вычислительной техники по получению новых данных по предмету исследования. Добавление соответствующего функционала позволяет использовать создаваемые программные продукты как инструмент для проведения новых исследований.

Начиная с 50 годов XX столетия лингвистами отдела диалектологии института языка, литературы и истории им. Г.Ибрагимова АН РТ (ИЯЛИ АН РТ) собирался и обрабатывался обширнейший материал по изучению территориального размещения носителей различных говоров татарского языка в регионах России. Результатом этих исследований явилось издание “Атласа татарских народных говоров Среднего Поволжья и Приуралья” в двух томах, вышедшего в 1989 году [1 - 3]. Издание атласа представляет собой итог многолетнего труда по исследованию говоров татарского языка методом лингвистической географии. Географический взгляд на особенности татарского языка позволяет объективно и на научной основе уточнить классификацию говоров, установить границы и пределы их распространения, уточнить ареал распространения многих важных с точки зрения языка явлений, в совокупности с историческими материалами раскрыть историю формирования большинства говоров и их носителей, показать на конкретном материале результаты многовековых связей народа с носителями родственных и неродственных языков. В настоящее время в ИЯЛИ АН РТ подготовлен к печати третий том атласа, охватывающего регионы Нижнего Поволжья и Сибири.

Представляя собой большой архив научных данных о языке, собранных по специально разработанной программе, атлас стал путеводной нитью для проведения многочисленных исследований по диалектологии татарского языка. Однако изучение атласа в формате 1989 года достаточно трудоемко и не позволяет в полной мере использовать опубликованный материал.

В 2011-2012 гг. при финансовой поддержке гранта РГНФ инициативной группой сотрудников института прикладной семиотики АН РТ, Казанского федерального

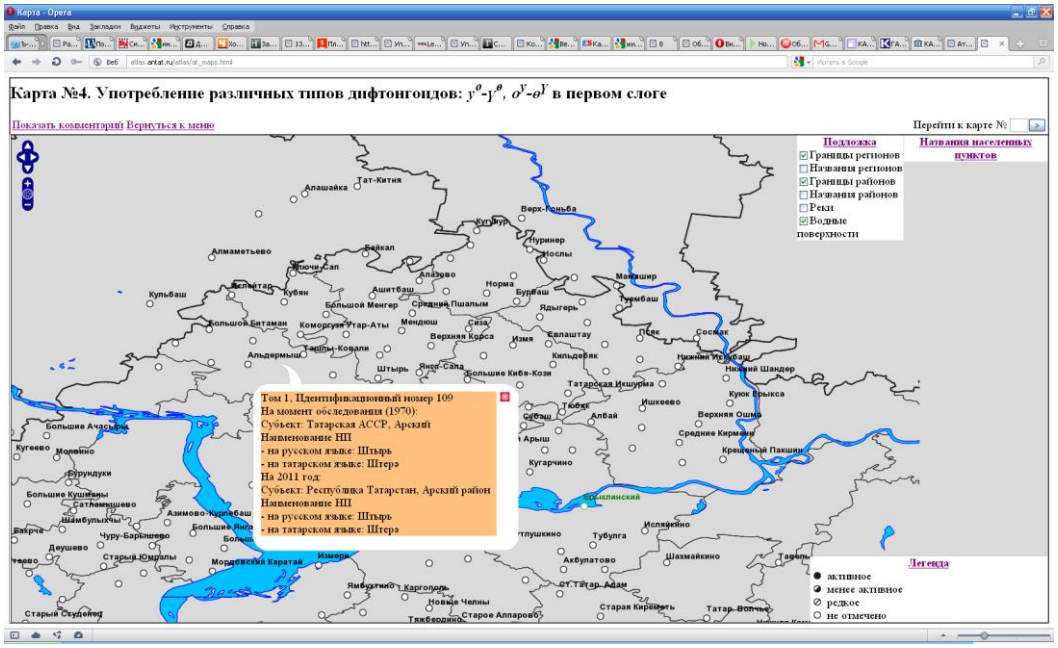
[1] Работа выполнена при финансовой поддержке РГНФ в рамках проекта «Электронный атлас татарских народных говоров» (проект №11-04-12020в)

университета совместно с лингвистами ИЯЛИ АН РТ была начата работа по созданию электронного атласа татарских говоров. Проект представляет собой Интернет-ресурс и включает в себя все материалы печатной версии атласа 1989 года, дополненные результатами неопубликованной части атласа. Электронный атлас состоит из картографической и атрибутивной баз данных. Картографическая база содержит пространственную информацию, разбитую на слои: административное деление РФ по состоянию на 1995 год, основные водные объекты на территории РФ, населенные пункты (НП), в которых проводился сбор информации. Разнесение информации по слоям позволяет проводить ее фильтрацию и показывать карты в виде удобном конечному пользователю. Атрибутивная база включает в себя данные по распределению языковых явлений по населенным пунктам, а также содержит дополнительную информацию по фонетическим, морфологическим и лексическим особенностям диалектов татарского языка.

В структурном отношении атлас состоит из 215 карт языковых явлений, исторических карт и сводных карт изоглосс, каждая из которых визуализируется в отдельной экранной форме. На картах отражены основные признаки диалектного различия, закономерности и характер распространения диалектных явлений. В атласе каждая карта снабжена комментариями, в которых приводится дополнительная информация по объектам атласа. Рабочий макет атласа выставлен в ИНТЕРНЕТ по адресу atlas.antat.ru.

Мы не будем в данной статье подробно останавливаться на преимуществах электронной версии. Сравнительный анализ печатной и электронной версии атласа дан в [4]. Целью настоящей статьи является демонстрация возможностей электронного варианта атласа для извлечения дополнительной информации о языковых явлениях в виде пользовательских запросов к базе данных. Хранение данных в виде электронной базы является особенно полезным для задач статистического анализа, который достаточно затруднительно проводить на книжной версии. Электронный атлас можно рассматривать как объект, заданный в трехмерном пространстве, где каждая карта, являясь плоскостью, несет в себе информацию по одному языковому явлению. Набор карт создает объемный образ, и запросы по выбранному набору позволяют сопоставить распределения соответствующих языковых явлений, что может оказаться полезным при решении задач кластеризации говоров.

База данных атласа содержит информацию о распределении языковых явлений по 1031 населенным пунктам (атрибутивная база данных), в которых проводились исследования, параллельно с этим в картографической базе данных содержится дополнительная информация о координатах этих НП. Каждую карту атласа можно классифицировать как ответ на запрос по данному языковому явлению. Кроме того ответы на некоторые простые запросы можно получить, обращаясь к различным объектам на карте. Например, двойной щелчок мышью на соответствующем населенном пункте идентифицирует этот населенный пункт и дает информацию о его местоположении и времени проведения обследования.



У пользователя также имеется возможность управлять визуализацией различных объектов на карте. В частности по желанию пользователя можно для идентификации населенного пункта использовать его татарское или русское наименование, или номер, использованный для обозначения в книжной версии атласа. Кроме того пользователь может управлять фильтрацией границ субъектов и районов, визуализацией названий субъектов и районов, а также визуализацией приведенных на карте водных поверхностей.

Описанные выше возможности представляют формы визуализации информации заложенной в базе данных. С другой стороны, рассматривая информацию, заложенную в базе данных как базовую, было бы полезным за счет запросов получать дополнительную аналитическую информацию, строить новые объекты на базе имеющихся. В атласе имеется список запросов, которые могут выбираться пользователем через экранные формы. Такие запросы сводятся к извлечению информации из картографической и атрибутивной баз данных и представлению этой информации в проекциях, которые интересны конечному пользователю.

Все запросы к базе к базе данных разбиты на два больших класса:

1. Запросы, которые выполняются на одной карте
2. Запросы, которые выполняются на отобранном пользователем наборе карт.

К первому классу относятся в основном запросы статистического характера, выполняемые в рамках одной карты. Поскольку все карты устроены одинаково, то такие запросы могут быть применены к каждой из 215 карт, составляющих содержание атласа. База данных атласа содержит информацию об административном делении РФ с локализацией населенных пунктов на уровне субъектов и районов РФ. Кроме того материал книжной версии атласа разбит на три тома: первые два тома охватывают татарские народные говоры, распространенные на территории областей и республик Среднего Поволжья и Приуралья; третий том - говоры татарского населения Западной Сибири. Это обстоятельство нашло отражение при построении запросов на картах атласа: с помощью экранных форм пользователь может выбирать необходимые параметры и тем самым проводить расчеты в рамках выбранного им территориального разбиения. Необходимо также отметить, что со времени проведения экспедиций некоторые населенные пункты перестали существовать или поменяли свои наименования. В связи с этим была проведена ревизия по всем населенным пунктам, включенным в атлас, и были выявлены НП, которые на сегодняшний день прекратили свое существование. На картах такие населенные пункты выводятся другим цветом.

Однако можно получить список «потерянных» НП выбрав соответствующий запрос. Ниже приведен список некоторых запросов, которые относятся к первому классу:

- Распределение количества языковых явлений в рамках одной карты относительно элемента выбранного территориального разбиения (заданного тома, субъекта РФ или района, в пределах области вокруг некоторого населенного пункта с заданным радиусом).
- Сравнительный анализ распределения суммарного количества исследованных населенных пунктов в пределах карты, тома, субъектов, районов
- Список населенных пунктов с заданным языковым явлением в рамках выбранного территориального разбиения
- Для фиксированного НП А вычислить расстояние до ближайшего населенного пункта с заданным языковым явлением: (с тем же, что и в НП А, языковым явлением, с отличным от НП А).
- Вычисление максимального (среднего) расстояния на множестве населенных пунктов, в которых наблюдается одно и то же языковое явление.
- Для выбранной пары языковых явлений вычисление максимального (среднего) расстояния на множестве населенных пунктов.
- Кластеризация населенных пунктов по языковым явлениям, построение изоглосс, характеризующих границы между языковыми явлениями
- Вычисление расстояний между двумя указанными населенными пунктами
- Составление списка исчезнувших НП в пределах некоторой территории.

Перечисленные запросы носят в основном статистический характер. Из приведенного списка следует, что каждый запрос является параметризированным и, на самом деле, включает в себя группу сходных запросов. Предполагается, что пользователь будет иметь возможность сам проводить разбиение карты на непересекающиеся области и собирать статистику в рамках выбранного разбиения. Например, пользователь может выбрать (закрепить) некоторый НП и, зафиксировав выбранные координаты в качестве начала координат, построить прямоугольную сетку с заданным шагом, и далее выполнить запрос с группировкой по построенной сетке. Такого рода запросы дают в руки пользователя возможность исследовать распределения языковых явлений в рамках выбранным им же самим шаблонам разбиения.

Несколько особняком стоит задача кластеризации (разбиения на карте на участки с определенными свойствами, например содержащими НП с одним языковым явлением) и построения изоглосс, которая сама по себе представляет достаточно серьезную проблему из области вычислительной геометрии[5]. Для решения этих задач могут быть использованы методы кластерного анализа[6,7], а также алгоритмы построения выпуклых оболочек. Однако конфигурация разброса языковых явлений по участкам карты может быть настолько сложной и причудливой, что предполагается автоматизировать только первый шаг в построении подобного разбиения, а доработку предоставить пользователю в ручном режиме. При этом для нанесения дополнительных границ на карту пользователь может использовать отдельный слой, который накладываясь на основную карту, даст необходимый результат. В атлас включены опытные образцы сводных карт, в которых при помощи изоглосс выделены и обобщены языковые явления, имеющие сходные очертания. Отметим, что составление сводных карт вручную является очень сложным и трудоемким процессом. Автоматизация процессов сравнения карт, описывающих распространение различных языковых явлений, позволило бы облегчить труд лингвистов и ускорить построений сводных карт.

Второй класс запросов составляют запросы по нескольким картам:

- Построение лингвистического паспорта для НП (набора НП) по всему набору карт (по выбранному подмножеству карт)
- Построение из базовых карт, представленных в атласе, новых карт, в которых визуализируется распределение некоторой совокупности языковых явлений по населенным пунктам.

- Сравнение распределений языковых явлений для различных карт

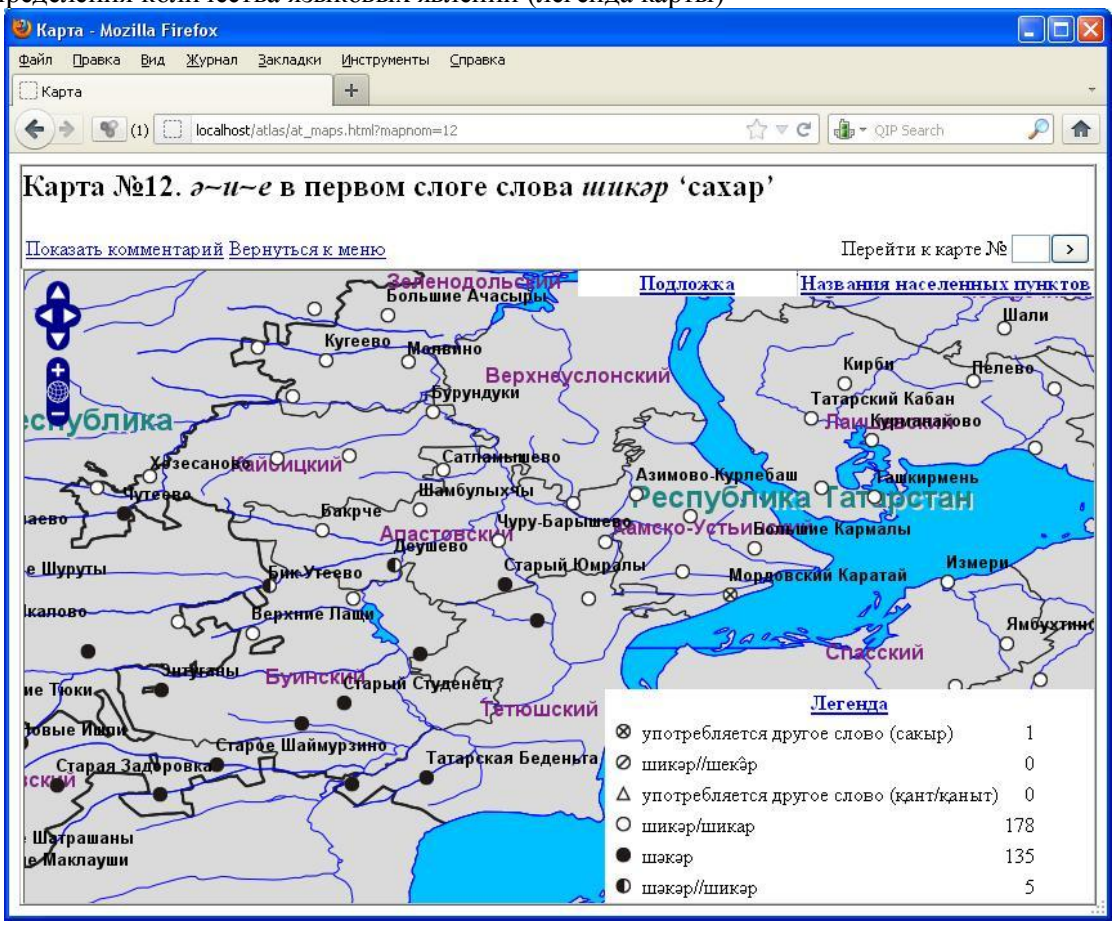
Под лингвистическим паспортом населенного пункта мы понимаем набор всех языковых явлений, приписанных этому поселению. Для диалектологов представляет интерес построение лингвистического паспорта для НП, с дальнейшим сравнением этих паспортов для разных НП (групп НП). При этом информация может быть сгруппирована по лексике, фонетике, морфологии и синтаксису. Такого рода запросы практически невозможно реализовать в ручном режиме и наличие подобных механизмов анализа информации представляет достаточно мощный инструмент для различных исследований.

Второй запрос определяет обратное отображение выбранного множества языковых явлений на населенные пункты. При этом набор языковых явлений, для который проводится исследование, может быть задан в виде логической формулы, что значительно расширяет возможности анализа.

Третий запрос представляет сравнительный анализ распределений различных языковых явлений по одной и той же территории. Это является важным для языковых явлений с разных позиций описывающих сходные параметры диалектов в языке (например, сравнения распределения определенных явлений фонетики и лексики). Эти запросы могут служить основой для проверки различных статистических гипотез на базе материалов атласа.

При проектировании запросов необходимо уметь отвечать на вопрос: как представлять результаты запроса? В большинстве случаев эти результаты могут быть представлены в виде отчетов. С другой стороны возникают задачи визуализации результатов запросов на картах атласа.

На рисунке ниже показан результат выполнения запроса по статистике распределения количества языковых явлений (легенда карты)



В данное время идет работа над реализацией некоторых типов запросов, изложенных в данной статье. По мере готовности результаты будут включены в электронную версию атласа.

Список литературы

1. Атлас татарских народных говоров Среднего Поволжья и Приуралья. / Под ред. Н. Б. Бургановой, Л. Т. Махмутовой, Ф. С. Баязитовой, Д. Б. Рамазановой, З. Р. Садыковой, Т. Х. Хайрутдиновой: в 2-х т.. - Казань, Татпрокаттехприбор, 1989 - 240 с.
2. Комментарии к Атласу татарских народных говоров среднего Поволжья и Приуралья. Казань, Татпрокаттехприбор, 1989. - 300 с.
3. Рамазанова Д. Б. Татарская диалектная лексикография и диалектная лексикология на рубеже III тысячелетия // Актуальные вопросы татарского языкознания. - Казань, 2002.- С. 101 - 106;
4. 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.
5. Препарата Ф., Шеймос М. Вычислительная геометрия: введение, М: Мир, 1989. - 478с.
6. Мандель И.Д. Кластерный анализ М.: Финансы и статистика, 1988. 176с.
7. Дюран Б., Одел П. Кластерный анализ М.: Статистика, 1977. 128с.