3rd International Conference on Computer Processing
in Turkic Languages (TURKLANG 2015)

PROJECT OF ELECTRONICAL ETNOLINGUISTIC TATAR DICTIONARY[1]

Farid Salimov, Rustem Salimov

Research Institute of Applied Semiotics of Tatarstan Academy of Sciences.
Kazan Federal University
Kazan, Tatarstan? Russia

Abstract

We analyze the first step of electronic ethno-linguistic resource creation (ethnoling.antat.ru), built on the basis of ethno-linguistic expeditions of the Institute of Language, Literature and Art Academy of Sciences of the Republic of Tatarstan. It was published more than 20 books and 300 scientific articles on the basis of systematization of materials collected by ethno-linguistic staff of IJALI in different years. Materials were collected in respect of ethno-cultural archaic dialect zones of Siberia, the Urals region, the Middle and Lower Volga region, densely inhabited by Tatar population.
The purpose of this project is to create an electronic resource which includes information in a structured way, extracted from major publications, based on the results of ethno-linguistic expeditions of IJALI AN RT. As a result of completion of the project electronic resource should be created and posted in the Internet. It should contain the terminology (ethno-linguistic) dictionaries with large amounts of live specimens of the Tatar language, collected in the expeditions. In addition, it is expected to bind created resources with electronic atlas of Tatar folk dialects.

• Введение

В современном мире наблюдается повышенный интерес к языку духовной народной культуры. Материалы живой диалектной речи и народной культурной традиции являются ценнейшим источником сведений о духовной культуре народа.

Начиная с 60-х годов XX столетия в институте языка, литературы и искусства им. Г. Ибрагимова Академии наук Республики Татарстан (ИЯЛИ АН РТ) проводится работа по сбору этнолингвистического материала по диалектам и говорам татар, проживающих в Республике Татарстан и других регионах России. Сбор этих данных проходил параллельно со сбором информационных материалов для атласа татарских народных говоров [Атлас татарских народных говоров Среднего Поволжья и Приуралья, 1989]. И хотя географические зоны полевых экспедиций во многом пересекались, программы научных исследований были различными, в результате лексический материал, собранный в рамках двух направлений отличался и дополнял друг друга. На основе анализа и систематизации собранных этнолингвистических материалов сотрудниками ИЯЛИ в различные годы было опубликовано более 20 монографий и около 300 научных статей. Материалы были собраны в архаически этнокультурном отношении диалектных зонах Сибири, регионах Урала, Среднего и Нижнего Поволжья, где компактно проживает татарское население.

• Электронный словарь

В 2011-2012 годах при финансовой поддержке РГНФ (проект №11-04-12020в) был создан электронный атлас татарских народных говоров (atlas.antat.ru), в котором на 215 картах отражены основные признаки диалектного различия для татарского языка, закономерности и характер распространения диалектных явлений в 28 регионах Российской Федерации [Салимов Ф.И., 2012]. Издание электронного атласа представляло первый опыт по интерпретации и представлению обширнейшего материала, изданного ИЯЛИ АН РТ в рамках печатного варианта атласа. Имеющиеся материалы были значительно дополнены дополнительными данными, собранными после 1989 года в

[1] Работа выполнена при финансовой поддержке РГНФ в рамках проекта «Создание электронного ресурса по этнолингвистическим (диалектно-фольклорным) материалам татарского языка» (проект № 14-04-12024)

сибирских регионах, а также библиотекой стандартных параметризованных запросов, позволяющих проводить многоаспектный анализ имеющегося материала с привязкой языковых явлений к географическим координатам.

Целью настоящего проекта является создание электронного ресурса, включающего в себя в структурированном виде информацию, извлеченную из основных публикаций, изданных по результатам этнолингвистических экспедиций ИЯЛИ АН РТ. После завершения проекта, в сети Интернет должен появиться электронный ресурс, в котором будут представлены терминологические (этнолингвистические) словари с большим объемом образцов живой татарской речи, собранных в полевых условиях. Кроме того предполагается привязать создаваемый ресурс к картам электронного атласа татарских говоров.

Основным источником для создания словаря были выбраны фундаментальные труды Ф.С. Баязитовой [Баязитова 2011, 2012], которые освещая в зависимости от разрабатываемой тематики различную диалектную лексику, имеют сходную конструкцию, позволяющую применять к анализу имеющегося материала похожие инструменты.

В этой статье описан опыт анализа научной монографии Ф.С. Баязитовой «Лексика татарской свадьбы (в контексте диалектных и фольклорных текстов)» /«ТУЙ ЙОЛАЛАРЫ ЛЕКСИКАСЫ (җирле сөйләш һәм фольклор текстлары яссылыгында)»/, изданной в 2011 году в Казани. В этой научной монографии на примерах различных образцов диалектных текстов исследуются исторические корни возникновения различных свадебных терминов, приводятся многочисленные примеры использования этих терминов в народной речи.

Монография представляет собой объемное научное исследование (около 960 страниц текста), которое включает большое количество иллюстративного материала, представляющего собой диалектные тексты в упрощенной транскрипции, записанные и расшифрованные с магнитных лент. Все приводимые примеры разнесены по темам, связанным с обрядовой свадебной лексикой, и сопровождаются семантическими толкованиями отдельных терминов. Диалектные лексические термины в большей своей части содержат ссылку на диалекты и говоры, к которых они встречаются.

Построение этнолингвистической базы данных предполагало выполнение определенной последовательности шагов и включало в себя следующие этапы:

1. Сканирование печатного экземпляра книги с целью создания ее электронного образа;
2. Анализ и систематизация имеющегося в книге материала;
3. Создание программы сегментации линейного текста с выделением набора необходимых фрагментов, вносимых в электронную базу данных; Автоматическая сегментация содержимого книги с последующей ручной проверкой и корректировкой выделенных сегментов;
4. Проектирование и создание базы данных, заполнение таблиц полученными данными;
5. Построение запросов к базе данных;
6. Создание клиентской части программы, размещение тестовой версии программы в Интернет.

1. Содержание книги Ф.С. Баязитовой представляет собой богатейший этнолингвистический источник, в котором с позиции различных диалектов татарского языка анализируется такая важная тематика, как свадебный обряд. Представленный в книге материал имеет ценность как в лингвистическом, так и этнографическом аспектах. Особую ценность представляет множество примеров практического использования терминов в диалектной речи татар. Поскольку у разработчиков в наличии имелся только печатный экземпляр книги, было проведено сканирование печатного материала с целью его перевода его в электронный образ с сохранением особенностей форматирования исходного текста. Эта работа с последующим ручным исправлением ошибок результатов сканирования потребовала достаточно больших усилий в виду большого объема исходного текста.

2. Книга Ф.С. Баязитовой состоит из многочисленных тематических групп (свадебные ритуалы, свадебные персонажи, сваты и сватовство, свадебная пища, свадебная одежда, и пр.), в каждом из которых описывается и систематизируется определенный набор свадебных терминов, относящийся к соответствующему разделу. В тексте книги перемешаны как сами термины, примеры их употребления в различных диалектах и говорах, семантическое толкование термина и другая информация. Основная задача анализа состояла в выявлении системы признаков, характеризующих различные фрагменты

текста: термины, ссылки на диалекты и говоры, в которых они употребляются, их семантическое описание, а также тексты примеров, которые служат для иллюстрации употребления соответствующих терминов в различных диалектах и говорах. Такой анализ представлял достаточно трудную задачу. Поскольку, изначально, книга не предполагала автоматическую обработку имеющегося материала, не всегда выдерживался единый язык разметки терминов, многие одинаковые в семантическом отношении фрагменты были оформлены различным образом, в приводимых словоформах не везде проставлены ссылки на диалекты и говоры, а имеющиеся ссылки – часто относились к различным по объему и содержанию текстовым фрагментам. В силу описанных причин, в определенных случаях, было довольно трудно автоматически, с помощью программы сегментации, определить на какой длины текстовый фрагмент распространяется приведенное значение параметра. Для выделения границ сегментов применялись различные подходы, в частности, использование форматов представления определенных фрагментов предложений в печатном варианте текста книги, анализ информации из некоторой окрестности анализируемого термина, построения словарей, содержащих наборы ключевых слов.

3. Была написана программа сегментации текста с выделением ключевых терминов, описания их семантики, указания диалекта или говора в которым встречается соответствующий термин, выделения примеров употребления термина в различных диалектах. Результаты работы программы сегментации представляются в виде набора специальных таблиц в формате текстового процессора WORD. При сегментации также учитывались расшифровки диалектизмов, которые встречаются в тексте книги и вынесены из текста в виде сносок. Эти диалектизмы составили отдельный словарь. К сожалению, вариативность представления информации в различных фрагментах текста оказалась достаточно большой, и несмотря на предпринятые усилия, не удалось полностью автоматизировать процесс сегментации. Поэтому,полученный на выходе программы размеченный текст дважды подвергался ручной обработке: сначала текст проверялся на ошибки определения границ выделенных фрагментов, далее откорректированный текст проверялся лингвистами на предмет его дополнения недостающими параметрами (в основном, указанием на диалекты и говоры, где употребляется тот или другой термин).

4. На основе отобранного материала был создан терминологический словарь диалектизмов (словоформ и словосочетаний). В состав словаря дополнительно была включена информация по транскрипции диалектизмов, а также перевод семантики включенных в словарь терминов на русский язык. При создании базы данных облегченная транскрипция, используемая Ф.С. Баязитовой была признана недостаточной и для каждого термина была предложена транслитерация терминов на письменные формы татарского языка с использованием символов Международного фонетического алфавита (МФА) [У.Ш.Байчура]. При этом преследовалась цель использования терминов словаря в диалектических корпусах. К настоящему времени общий объем словаря составляет около 3500 словоформ и словосочетаний.

В базу данных была включена обширная коллекция диалектных текстов, собранных в полевых экспедициях. Элементы базового словаря и примеры связаны между собой отношением один ко многим, которое позволяет по заданному термину найти набор примеров по употреблению данного термина в различных диалектах. Была разработана специальная программа согласованной загрузки данных из обработанных текстовых фрагментов. Дополнительно, данные базового словаря были индексированы специальными пометками, что позволило строить запросы по набору словоформ, относящихся к отдельным разделам свадебной терминологии. В состав базы данных также были включены дополнительные словари по диалектизмам татарского языка, которые непосредственно не относятся к свадебной тематике, но встречаются в тексте книги и расшифровываются в виде сносок. Дополнительно в базу данных включена информация по населенным пунктам, где собиралась информация, по информантам, по научным источникам, которые были использованы при создании книги. При анализе списка населенных пунктов добавлена информация по диалектам и говорам, характеризующим данный населенный пункт, эта информация сравнивалась с подобной информацией для населенных пунктов, описанных в составе атласа татарских говоров. Следует отметить, что из более чем 200 населенных пунктов, где собиралась информация по свадебной лексике, только лишь половина описана в составе атласа. Это объясняется тем, что информация по этнолигвистическим материалам в основном собиралась в отдаленных, малонаселенных селах, в которых в силу их

удаленности от "цивилизации", сохранились сведения о первобытном языке и культуре. Однако это обстоятельство создало определенные сложности при определении географических координат соответствующих населенных пунктов. Со времени проведения экспедиций, многие населенные пункты прекратили свое существование, и было довольно трудно найти по ним необходимую информацию.

5. Во процессе аботы над проектом создана библиотека типовых запросов по следующим направлениям:

• Запросы, которые по выбранной словоформе (словосочетанию) и диалекту (говору) показывают на экране наборы образцов текстов, иллюстрирующих использование этой словоформы (словосочетания) на примерах живой народной речи.

• Построения списка словоформ, связанных с некоторой темой свадебного обряда.

• Запросы по формированию карт, которые показывают географию распространения соответствующей словоформы. Поскольку в книжной версии не указан конкретный список населенных пунктов, связанных с распространением того или иного термина, при построении карт за основу было взято разбиение населенных пунктов по диалектам и говорам, зафиксированное в электронном атласе народных говоров. Такое представление, не являясь абсолютно точным, все же дает определенную картину о географическом распространении лексических терминов, использованных при написании книги.

• В настоящее время производится индексация базового словаря, которая позволит строить запросы для выявления терминов-синонимов, употребляемых в различных диалектах.

6. Программа создается как интернет-ресурс. Тестовый вариант программы доступен по адресу ethnoling.antat.ru. Основную часть программы составляет базовый словарь терминов, который позволяет по заданному термину строить множество примеров, в которых встречается употребление этого термина. Визуализация содержимого словаря в программе может производиться в различных режимах: в алфавитном порядке, в тематическом плане. Выбор режима просмотра определяется конечным пользователем. В программе реализовано два режима работы: режим пользователя, и режим администратора системы. Режим администратора необходим для модификации структуры базы данных и заполнения ее содержимого новыми данными.

- **Заключение**

Представленный в базе данных материал может служить ресурсом для создания диалектологических подкорпусов татарского языка, этнолингвистических словарей, в частности построения словаря диалектизмов.

В будущем, предполагается пополнение словаря новыми словоформами, которые позволят точнее дифференцировать диалекты татарского языка, полнее описывать их особенности. Отметим, что в словарь вошло много словоформ и словосочетаний, которые отсутствуют в Большом диалектологическом словаре татарского языка 2009 года издания.

Литература

1. У.Ш.Байчура "Звуковой строй татарского языка", изд-во КГУ, 1959г.

2. Атлас татарских народных говоров Среднего Поволжья и Приуралья / науч. редакторы: Н. Б. Бурганова, Л. Т. Махмутова; Составители: Н. Б. Бурганова, Л. Т. Махмутова (I т.), Ф. С. Баязитова, Д. Б. Рамазанова, 3. Р. Садыкова, Т. Х. Хайрутдинова (2 т.).— Казань, 1989. Прил.: Комментарии к Атласу. — Казань, 1989. — 300 с.

3. Баязитова Ф.С. (2011) Туй йолалары лексикасы (җирле сөйләш һәм фольклор текстлары яссылыгында). – Казан: Паравитта, 2011. – 976 б.

4.Баязитова Ф.С. (2012)Халык традицияләре лексикасы: бишек туе (йола һәм фольклор текстлары яссылыгында). – Казан: Алма-Лит, – 331 б.

5 Салимов Ф.И., Рамазанова Д.Б., Пилюгин А.Г., Салимов Р.Ф. (2012) Электронная версия атласа татарских народных говоров// Вестник татарского государственно-гуманитарного педагогического университета, Казань, с.205-210

6. Салимов Ф.И. Пилюгин Г.А. Ершов С.А. (2012) Электронный атлас татарских народных говоров как инструмент исследования - Труды татарской школы по компьютерной и когнитивной лингвистике, TEL-2012, ФЭН, АН РТ, Казань, с.48-54