УДК 004.912

ЭТНОЛИНГВИСТИЧЕСКИЙ ЭЛЕКТРОННЫЙ СЛОВАРЬ ТЕРМИНОВ ТАТАРСКОГО ЯЗЫКА

Ф.И Салимов., Р.Ф. Салимов
Казанский федеральный университет, Казань
Farid.Salimov@kpfu.ru, Rust1k@gmail.com

В статье описан опыт создания электронного этнолингвистического словаря, построенного на базе материалов, собранных учеными ИЯЛИ АН РТ во время полевых экспедиций.

Ключевые слова: *этнолингвистика, диалекты, татарский язык, базы данных*

Благодарности: Работа выполнена при финансовой поддержке РГНФ (проект № 14-04-12024)

При создании лингвистических электронных ресурсов большое значение имеет коллективный опыт исследований различных архивных материалов, собранных в течение длительного времени, и опубликованный в виде традиционных форм представления информации: книг, различного рода научных публикаций. Конечно, идеальным вариантом при создании электронных баз данных представляется анализ и размещение первичных источников информации, которые хранятся в виде бумажных картотек и в сжатом виде представляют экспериментальные данные, собранные коллективами ученых в различных лингвистических экспедициях. К сожалению, такие архивы по разным причинам являются малодоступными, подвержены различным внешним воздействиям, имеют высокую вероятность разрушения. В качестве таких примеров можно упомянуть картотеку данных русского диалектологического атласа (ДАРЯ), которая в большей своей части была утеряна в результате проведения ремонтных работ в институте Русского языка [1]. Подобная участь постигла картотеку атласа татарских народных говоров [2]. Одной из причин такого положения дел является падение интереса к сохранности исходных данных после их первичного анализа и выхода в свет книжных публикаций, в которых отражены результаты научных исследований.

При отсутствии первичных архивов приходится опираться на вторичные источники в виде изданных научных трудов, в которых исходные данные подвергнуты определенному анализу. Несмотря на то, что такой подход является менее информативным и содержит определенную субъективную окраску исследователей, он имеет и некоторые положительные моменты, поскольку при отборе материала могут быть учтены результаты его первичной обработки.

Начиная с конца 50-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. Материалы собираются в архаически этнокультурном отношении диалектных зонах Сибири, регионах Урала, Среднего и Нижнего Поволжья. На основе анализа и систематизации собранных данных сотрудниками ИЯЛИ в различные годы было опубликовано более 20 монографий и около 300 научных статей.

В данной статье описан опыт по созданию электронного этнолингвистического терминологического словаря, включающего в себя в структурированном виде информацию, извлеченную из основных публикаций, изданных по результатам этнолингвистических экспедиций ИЯЛИ АН РТ. Источниками для словаря были выбраны объемные фундаментальные научные труды Баязитовой Ф.С., составляющие семейный цикл [3,4,5]. Тематика этих книг включает в себя описание лексики, обычаев, обрядов, связанных с рождением ребенка, свадебной церемонией, со смертью и погребением. Все книги Баязитовой изданы на татарском языке, имеют сходное строение, их содержание построено на тематико-гнездовом принципе, для большинства терминов дано их семантическое описание. Кроме того, каждая книга содержит большое количество образцов живой речи татар, которые приведены в качестве иллюстративного материала со ссылкой на диалекты и говоры, в которых они встречаются. Это

обстоятельство, безусловно, повышает информативность публикаций и позволяет использовать их в виде источников для создания электронных информационных ресурсов.

Создание татарского этнолингвистического словаря на основе опубликованных печатных материалов состояло из нескольких этапов и включало в себя решение ряда задач:

1. Сканирование печатного материала книг с целью получения электронного образа;
2. Анализ и систематизация имеющихся в книгах материалов;
3. Сегментация содержимого книг с целью создания фрагментов, которые составляют содержание статей электронного словаря;
4. Заполнение базы данных, создание библиотеки запросов;
5. Создание клиентской части программы с размещением ее в сети Интернет.

1. Реализация первого этапа носила в основном технический характер. Поскольку предполагалась автоматическая обработка полученного электронного образа, то особое внимание уделялось выбору формата хранения электронных данных, Одна из основных задач при конвертировании состояла в стремлении максимально сохранить стиль оформления текста сканируемого материала с целью его дальнейшего использования процедурами сегментации текста. Наиболее подходящим оказался формат текстового процессора WORD. Поскольку существующие сканеры не всегда справлялись с поставленной задачей, потребовался просмотр и корректировка результатов сканирования в ручном режиме. С учетом объема исходного материала эта часто работы была достаточно затратной и потребовала больших усилий.

2. Книги Ф.С. Баязитовой имеют определенную структуру: любая книга состоит из многочисленных тематических групп (гнезд), в каждой из которых исследуется и систематизируется определенный набор терминов, относящийся к разделу раскрываемой темы. Содержание гнезда характеризует определенные ритуалы, связанные с рождением ребенка; описание лиц, принимающих участие в совершении обряда; описание предметов, которые используются при выполнении обрядовых действий; сами обрядовые действия; поверья, свадебные или погребальные ритуалы, характеристики персонажей, принимающих участие в церемониях, свадебная пища, свадебная одежда, и пр. Каждое такое гнездо занимает определенный фрагмент в тексте книги. При этом границы соответствующих фрагментов порой сильно размыты и явно не выделены. Основная задача анализа текста книги состояла в выявлении системы признаков, характеризующих различные фрагменты текста с дальнейшим автоматическим выделением частей текста, относящихся к определенному термину. При анализе выделялся сам термин, его семантическое описание при его наличии в тексте, ссылки на диалекты и говоры, в которых этот термин используется, множество примеров употребления термина в различных диалектах.

3. Описание семантики терминов, приводимых в книгах Баязитовой, даже для опытных лингвистов представляет достаточно непростую задачу. Известно, что в этнолингвистических терминах отражается определенная «картина мира», которая формируется в этносе; в терминах прослеживается связь языка с элементами народной культуры, всех ее жанров и форм. Поэтому содержание того или иного термина можно определить только в контексте определенных явлений, событий. Тем не менее, в рамках проекта по созданию терминологического словаря с учетом объема опубликованного материала была предпринята попытка предварительной автоматической обработки имеющихся текстов с последующей ручной корректировкой границ выделенных фрагментов. Была создана процедура сегментации текста с выделением ключевых терминов, фрагментов, описывающих семантику, выделения примеров употребления термина в различных диалектах. При разбиении текста на фрагменты процедура ориентировалась в основном на формат оформления соответствующих частей текста-источника, с дополнительным анализом материала на присутствие некоторых ключевых слов, приведенных в справочниках. Этот подход позволял лишь приблизительно определять границы выделяемых фрагментов. Причина такого положения дел в первую очередь состояла в различии стилевого оформления одинаковых по смыслу фрагментов даже в пределах одного источника, не всегда

поддерживался единый язык разметки различных частей текста, нередко в приводимых примерах информация по диалектам и говорам была неполной. При написании книг не предполагалась обработка текстов в автоматическом режиме, они, прежде всего, были написаны для специалистов. Поэтому после программного выделения фрагментов, точность определения их границ проверялась лингвистом, при обнаружении ошибок производилась ручная корректировка границ. Дополнительно к основному словарю терминов был построен словарь диалектизмов, которые встречаются в тексте книг, но вынесены из текста в виде сносок. Ниже приведен фрагмент промежуточной рабочей таблицы, которая сформирована в результате работы процедуры сегментирования

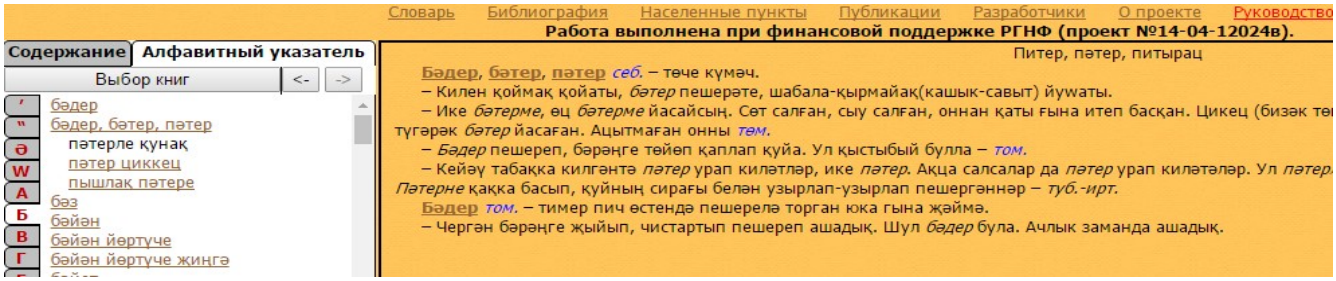
3	себ.	**Айшы**	**Айшы** *себ.* – гыйшык.
3	себ.	**Айшы уты**	**Айшы уты** *себ.* – гыйшык, мәхәббәт уты.
4	т м.		Қапқадин қарадым сәне, *Айшы уты* йандыра мәне – *т м.*
3	миш. д.	**Айшиклык йырлары**	**Айшиклык** [гар. айшик + афф. -лык] **йырлары** *миш.д.* – гыйшық, мәхәббәт җырлары. **Ашик тоткан кызлар** *чпр.* – гыйшык тоткан кызлар.
4	чпр.		– Кич утырганда *ашиклык йырлары* йырлыйбыз, кулга йаулык тотып. *Ашик тоткан кызлар* бик хуwат йырлыйлар ашиклык йырла--- рын– *чпр.*
4	к рш.		Әти дә кәбәмнең миленчасы Сәтенә йез пыт он тартадыр. Синдәй матурымны күргән сайын Йандин ашиклыгым артадыр – *к рш.*

В этой таблице третий столбец определяет выделяемый термин, второй столбец характеризует диалект, четвертый столбец содержит или семантическое описание термина, или примеры употребления термина в живой речи (характер записанной в строке информации определяется кодом в первом столбце). Зеленым цветом в тексте выделены диалектизмы, которые носят вспомогательный характер и расшифрованы в сносках книг. Эти термины были выделены программой в отдельный словарь.

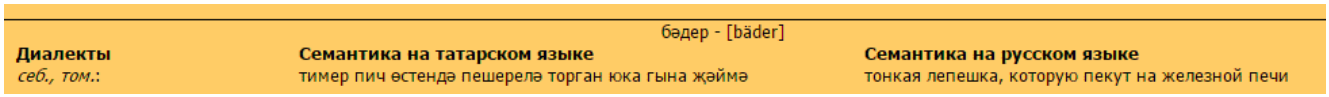
4. В результате работы была создана база данных терминологического словаря, включающего в себя этнолингвистические термины, их описания, многочисленные примеры употребления терминов в живой речи татар с указанием на диалекты, в которых они встречаются. Словарь был дополнен информацией по фонетической транскрипции диалектизмов, а также для части терминов словаря, которые относятся к родильной и свадебной тематики был осуществлен перевод семантики на русский язык. При создании базы данных используемая в книгах Ф.С. Баязитовой облегченная транскрипция, была признана недостаточной: для каждого термина была предложена транслитерация терминов на письменные формы татарского языка с использованием символов Международного фонетического алфавита (МФА) [6]. При этом преследовалась цель использования терминов словаря в диалектических корпусах. К настоящему времени общий объем словаря составляет около 6000 словоформ и словосочетаний.

Дополнительно в базу данных включена информация по населенным пунктам, где собиралась информация, по информантам, по научным источникам, которые были использованы при создании книги. Кроме того, была реализована связь между терминов построенного словаря с картами атласа татарских народных говоров [7]. Такой результат, не давая точной картины географии распространения термина, используя диалектное членение татарского языка, позволяет приблизительно определить населенные пункты, где может употребляться данный термин.

5. Программа размещена в сети Интернет по адресу ethnoling.antat.ru. Ниже на рисунке показан образ основного экрана программы



В левой части экрана визуализируется список терминов, начинающихся на определенный символ, выбираемый пользователем При этом пользователь также может выбирать источник (книгу), может выбирать способ представления информации (тематический в виде определенных гнезд или алфавитный). В правой части экрана показывается подробная информация по выбранному термину (его семантика, примеры употребления в языке, указания на диалекты, в которых этот термин употребляется). Названия диалектов выделены синим цветом. Выбор диалекта мышью позволяет активизировать карту регионов РФ с визуализацией географического расположения населенных пунктов, в которых распространен указанный диалект. В нижней части экрана приведены фонетическая транскрипция термина, его семантическое описание на татарском и русском языках:



Создаваемый терминологический словарь имеет самостоятельное значение и может быть использован при обучении татарскому языку. Кроме того большой набор примеров употребления различных терминов в живой речи материал может служить ресурсом для создания диалектологических подкорпусов татарского языка, этнолингвистических словарей.

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. Атлас татарских народных говоров. 2-е изд. – доп./ Под ред. Д. Б. Рамазановой, Т. Х. Хайрутдиновой - Казань, ИЯЛИ, 2015 - 631 с.
3. Баязитова Ф.С. Туй йолалары лексикасы (җырле сөйләш һәм фольклор текстлары яссылыгында). – Казан: Паравитта, 2011. – 976 б.
4. Баязитова Ф.С. Халык традицияләре лексикасы: бишек туе (йола һәм фольклор текстлары яссылыгында). – Казан: Алма-Лит, 2012 – 331 б.
5. Баязитова Ф.С. Халык традицияләре лексикасы: соңгы туй (дини фольклор һәм җырле сөйләш текстлары яссылыгында). – Казань, 2015. – 710 б.
6. У.Ш.Байчура "Звуковой строй татарского языка", изд-во КГУ, 1959г., Ч. 1. 183 с.
7. 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

ETHNOLINGUISTIC ELECTRONIC DICTIONARY TERMS TATAR LANGUAGE

F.I. Salimov, R.F.Salimov
Kazan Federal University, Kazan
Farid.Salimov@kpfu.ru, Rust1k@gmail.com

The article describes the experience of creating an electronic dictionary of ethno-linguistic, built on the basis of materials collected by scientists of the Institute of language, literature and art of Tatarstan Academy of Sciences during field expeditions.

Keywords: *dialects of the Russian language, database*