УДК 419. 992 61
С 42

З.А. СИРАЗИТДИНОВ
кандидат филологических наук
зав. лабораторией лингвистики и информационных технологий, ИИЯЛ УНЦ РАН

ИНФОРМАЦИОННЫЕ РЕСУРСЫ БАШКИРСКОГО ЯЗЫКА КАК ИСТОЧНИКИ ДЛЯ СРАВНИТЕЛЬНЫХ ИССЛЕДОВАНИЙ ТЮРКСКИХ ЯЗЫКОВ

Статья посвящена проблеме использования лингвистических информационных ресурсов в сравнительных исследованиях тюркских языков. Автором рассматривается детально ресурс компьютерного (машинного) фонда башкирского языка, определяются возможности применения в сравнительных и сопоставительных исследованиях.

Ключевые слова: информатизация, Машинный фонд башкирского языка, лексикографический подфонд, грамматический подфонд.

Стремительная информатизация человеческого общества вовлекает в свою орбиту и лингвистическую сферу деятельности. На сегодня многие тюркские языки переводят свои ресурсы в электронный вариант: появляются компьютерные словари, создаются базы филологических данных и целые информационные ресурсы с поисковыми системами как на компакт дисках, так и в сети Интернет [1]. Не остается в стороне от рассматриваемого процесса и язык башкирского народа [2]. Вовлечение национального языка в информационную сферу открывает широкие перспективы для сравнительных исследований лингвистических реалий родственных языков, одним из объектов которых выступает башкирский язык. Эти перспективы связаны с тем, что открывается быстрый доступ к материалам исследования большого объема. Однако, есть и сдерживающие факторы в использовании информационных ресурсов. Таковыми являются:

1) достоверность лингвистических данных;

2) компетентность составителей ресурсов;

3) время функционирования сайта с данным ресурсом в сети Интернет.

Эти факторы касаются, в основном, информационных данных википедии, коммерческих организаций и частных лиц. Но такие проблемы не возникают для лингвистических ресурсов, созданных в академических структурах или научно-исследовательских институтах авторитетных вузов, имеющих на начальной странице или вкладышах компакт дисков наименование создателя ресурса, а базы данных снабжены ссылками на печатные первоисточники. Одним из таких серьезных ресурсов по

башкирскому языку является Машинный фонд башкирского языка (МФБЯ), созданная в лаборатории лингвистики и информационных технологий Института истории, языка и литературы Уфимского научного центра Российской академии наук [3].

Базы данных МФБЯ построены на основе изданных печатных продукциий, с указанием источников, за достоверность представленных данных лаборатория несет ответственность. Все существующие базы имеют свои поисковые аппараты, меню пользователя представлен на трех языках: башкирском, русском и английском.

Идеи создания академических лингвистических компьютерных баз данных исследовательского профиля в виде фондов по тюркским языкам выдвигались в 80-х годах прошлого столетия [4]. В те же годы Пиотровским Р.Г., Щербой А.М. и Гузевым В.Г. поднимался вопрос и о разработке единого машинного фонда тюркских языков [5]. По последнему фонду даже был создан координационный совет в г. Чимкенте во главе с известным ученым в области математической и прикладной лингвистики профессором К.Б.Бектаевым. Катализатором этих идей было начало создания Машинного фонда русского языка (МФРЯ) [6]. Отметим, что МФРЯ сейчас активно действует и пользуется авторитетом в научных кругах [7].

Но к сожалению, на сегодня в тюркском мире такой ресурс создан только для башкирского языка, на котором мы хотели бы остановиться в нашей статье.

Основные положения разработки МФБЯ были выдвинуты нами еще в начале 90-х годов [8]. Непосредственная работа по созданию фонда началась в конце 2003 г. с разработки концепции и структуры фонда как системы баз данных, работающей на отдельных компьютерах Института истории, языка и литературы УНЦ РАН. В дальнейшем было принято решение о разработке сетевой концепции для ИИЯЛ и гуманитарных вузов. В 2005 г. окончательно утвердилась концепция машинного фонда в виде открытой сетевой системы с доступом через Интернет.

За эти годы фонд уже принял устоявшуюся стабильную структурную форму, хотя говорить о завершении работы над МФБЯ еще рано.

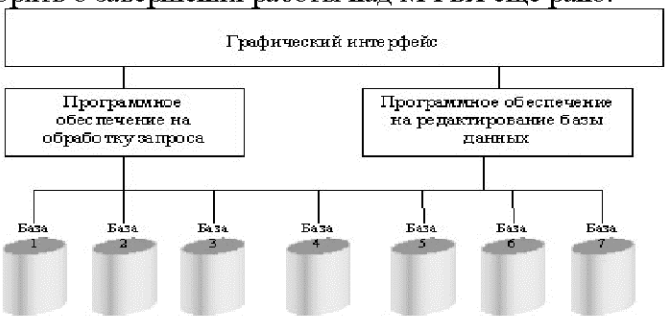

Рис. 1. Структура машинного фонда башкирского языка.

Что же собой представляет этот фонд, какие возможности в ней заложены? Машинный фонд башкирского языка — это специально разработанная для ученых-лингвистов, преподавателей, аспирантов и студентов система поиска филологической информации со своими базыми данных. Причем, как было отмечено выше, базы данных построены на основе изданных печатных продукций и имеют соответствующие ссылки. Данная информационная система включает в себя специализированные базы данных, интерфейсы для выполнения запросов к базам данных и программные обеспечения для обработки этих запросов. На рис. 1 представлена структурная схема информационной системы.

Фонд разработан на основе системы управления базами данных Oracle и функционирует на специально выделенном сервере по адресу www.mfbl.ru.

На сегодня эта система имеет 7 крупных баз данных, которые образуют подфонды единого машинного фонда:

- подфонд генеральной картотеки;
- лексикографический подфонд;
- грамматический подфонд;
- подфонд каталога рукописных книг;
- подфонд каталога старопечатных книг;
- экспериментально-фонетический подфонд;
- диалектологический подфонд.

1. **Подфонд генеральной картотеки** содержит основную информацию о лексической системе языка. В базе в структурированном виде находятся более 100000 корнеслов и производных всех пластов языка. Структура представления данных в генеральной картотеке представлена на рис. 2.

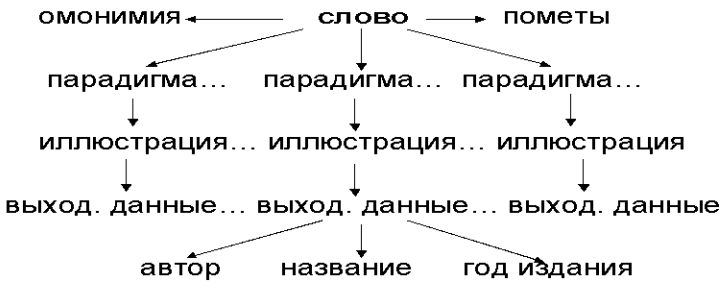

Рис. 2. Структура представления данных в генеральной картотеке.

Подфонд дает информацию о таких характеристиках слова как часть речи, происхождение слова, стиль, диалектное/литературное, историзм/ архаизм/неологизм, нарицательное/собственное и др. Разработаны более 50 помет для каждой лексической единицы. Каждое слово представлено в базе своими парадигмами употребления в языке, и для каждой парадигмы приводятся примеры употребления из художественной, публицистической и научной литературы, устного народного творчества с указанием источника печатного материала (названия, автора и даты издания). На данном этапе примеры употребления вводятся из картотеки отдела языкознания Института истории, языка и литературы УНЦ РАН.

База данных этого подфонда связана с базами других подфондов, в частности, можно определить в каких лексикографических работах (это уже подфонд лексикографический) зафиксировано то или иное слово из генеральной картотеки.

2. **Лексикографический подфонд** дает информацию о слове исходя из существующих словарей башкирского языка. База данных этого подфонда на сегодня состоит из 62-х структурированных словарей с общим объемом словарных статей порядка 500 000 единиц.

В базе представлены академические и учебные словари: одноязычные и двуязычные, частотные, терминологические, фразеологические, синонимические; словари-справочники (названий населенных пунктов, водных объектов, названий улиц городов, названий горных объектов РБ) и др. На рис. 3. показан вид интерфейса лексикографического подфонда.

В рассматриваемом подфонде пользователь может производить поиск либо по начальному фрагменту слова, либо по любой его части. Найденную словарную статью можно забрать для дальнейшей работы.

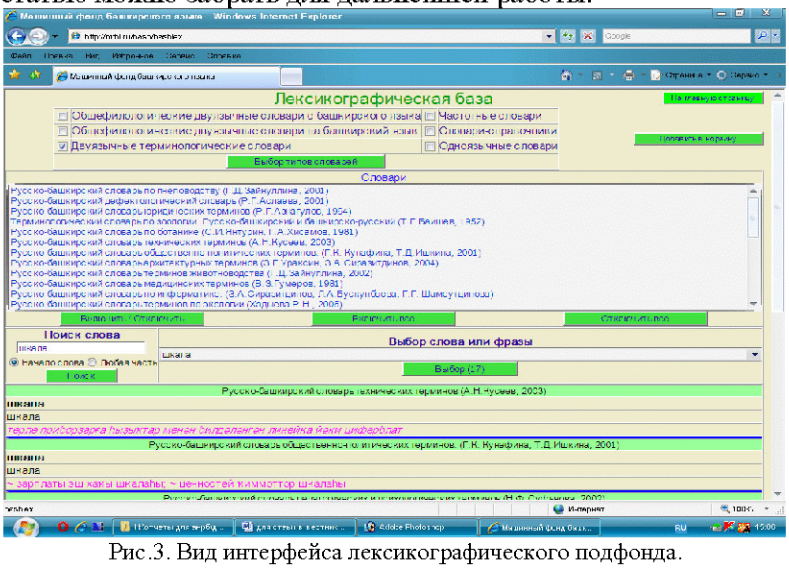

Рис.3. Вид интерфейса лексикографического подфонда.

3. В экспериментально-фонетическом подфонде даются артикуляционные и экспериментальные характеристики вокализмов и консонант башкирского языка. На данный момент экспериментальные характеристики представлены амплитудно-частотными графиками и сонограммами, которые получены сотрудниками лаборатории. Фонемы можно прослушать. В подфонде также представлен фонетический словарь объемом в 8000 единиц. Кроме научного интереса, фонетический подфонд представляет большой интерес для изучающих башкирский язык самостоятельно.

4. С целью информирования населения и научного мира о существующих источниках башкирского письменного литературного языка в МФБЯ представлены два каталога: рукописных и старопечатных книг. Общий объем баз этих каталогов составляет более 3000 единиц. Каталог рукописных книг содержит в себе описание рукописных книг из фонда Института истории, языка и литературы УНЦ РАН.

Описываются следующие характеристики книг: заглавие (перевод на русском языке), заглавие в транслитерации, автор (автор произведения), автор в транслитерации, переписчик (кем переписана рукописная книга), год (дата, когда переписана книга переписчиком), объем (в листах), формат (кодекса и текста, количество строк на странице), характеристика, аннотация (краткое описание содержания книги), коллекция (кем, где и когда найдена книга, последний владелец книги), язык (язык текста: арабский, старотюркский, османский и т.д.), палеографические сведения (бумага, чернила, почерк, пагинация, наличие комментариев), библиография (ссылки на данную рукописную книгу), сигл (где хранится рукописная книга), шифр (шифр книги в хранилище).

Пользователю предоставляется возможность производить поиск книги по 6 полям: заглавию, языку, автору, году, сиглу и шифру. Для того, чтобы не нагромождать экран информацией, реализована возможность выбора отображаемой информации на выходе.

Каталог старопечатных книг содержит описание арабографичных книг, хранящихся в фонде Национального музея РБ. Книги описаны по 15 полям: автор, заглавие книги, издатель книги, место издания книги, типография (название типографии), год издания, часть (если книга издавалась в нескольких частях), объем (в страницах), формат книги, переплет (наличие или отсутствие переплета, вид переплета), характеристика (титульный лист, сигнатура, пагинация, колонцифры), аннотация (краткое описание содержания книги), языка (язык текста книги), сигл (место хранения книги), шифр (шифр хранения в хранилище). Предоставляется возможность производить поиск книги по 9 полям: заглавию, языку, автору, году, сиглу, шифру, месту издания, издателю, типографии.

5. Работа по диалектологическому подфонду была начата в 2008 году. На сегодня этот подфонд состоит из трех самостоятельных баз данных. Это лексическая, картографическая и текстологическая базы.

В лексическом разделе представлена информация о диалектной лексике. База данных разработана на основе словарей диалектов башкирского языка и содержит более 52000 диалектных единиц с шестью информационными полями: само диалектное слово, часть речи, диалект, говор, литературная норма, русский перевод [9]. По каждому из этих полей осуществляется гибкий поиск информации.

В картографическом разделе представлена информация по изданному диалектологическому атласу башкирского языка [10]. База данных диалектологического атласа позволяет выбрать типы языковых явлений (фонетический, морфологический, синтаксический, лексический), для каждого типа определены конкретные изоглоссы. Изоглоссы выделяются по 250 опорным пунктам республики и сопредельных районов, по которым сотрудниками Института истории, языка и литературы были собраны экспедиционные материалы с 1973 г. по 1980 г.

Текстологическая база данных начала создаваться в 2009 г. Здесь представлены образцы устной речи, собранные во время экспедиций в разные годы сотрудниками ИИЯЛ. В базе можно производить выбор текстов по лингвистическим и экстралингвистическим параметрам: диалект, говор, год записи, образование информатора, пол информатора, возраст информатора, национальность.

6. Грамматический подфонд включает:

а) гипертекстовые представления академической грамматики [11];

б) алгоритмическое представление словоизменительной системы башкирского языка, являющейся основой нашего монографического исследования [12];

в) морфологические данные по именным и глагольным формам башкирского языка, включающие информацию по структурным моделям словоизменения и формообразования, количественные данные о реализации этих форм в речи. База данных реализована на основе выборок из текстов трех функциональных стилей языка: публицистика, проза и наука. Общий объем базы составляет 1000000 словоформ.

Интерфейс морфологической базы позволяет производить как поиск употребления отдельных морфологических категорий, так и конкретных аллофонов аффиксов этих категорий. Предусмотрена возможность поиска по сочетаниям нескольких категорий и конкретных аффиксов этих категорий.

Машинный фонд башкирского языка не только предоставляет доступ к языковым источникам, но также является инструментом для лингвистических исследований, поскольку фонд в самом начале

планировался именно в этом русле. Проиллюстрируем это на следующих примерах.

Например, языковеда интересует наличие формантов –**балҡан** или -**ғазы/-гәзе** в топонимах Башкортостана. Поиск по маске в базе топонимических данных сразу же выдает все топонимы с данными формантами с указанием типа объекта и места нахождения. За короткое время можно найти десятки, сотни, тысячи разных примеров.

Если необходимо выявить наличие диалектных слов, имеющих значение “пуговица”, то лексикографическая база диалектного подфонда по полю “русский перевод” выдает 63 диалектных единиц.

Машинный фонд окрывает большие возможности и для научных исследований в области морфологии языка, поскольку иллюстрационные материалы даются по морфологическим парадигмам слов. В генеральной картотеке можно сразу же просмотреть тексты на данную грамматическую форму и забрать для дальнейшей работы.

Мы надеемся, что МФБЯ будет активно использоваться тюркологами в сравнительно-исторических исследованиях.

ЛИТЕРАТУРА

1. http://sozdik.kz; http://solver.uz; http://www. muhranoff.travel.ru/turslov.htm; http://tili.kg/dict; maturtel.ru; http://abc.marlamuter.ru/index.php/term/rus-tat,2844-bol-shoij. xhtml; Электронный словарь ABBYY Lingvo x5; http:// www.medialingua.ru; http://til.gov.kz/; http://mfbl.ru; http://fodor.ii.metu.edu.tr/ content/metu-turkish-corpus; http://mfbl.ru/bashkorp/korpus;http://web-corpora.net/ Tatar
2. Corpus/search/?interface_language=ru.
3. http://blang.ru; http://www.mfbl.ru; huzlek.bashqort.com; http://ru.glosbe. com/ba/ru; Словари башкирского языка. ООО “Корал”, Уфа, 2005.
4. Адрес Машинного фонда башкирского языка http://www.mfbl.ru
5. *Галиуллин К.Р.* Машинный фонд татарского языка: особенности формирования и функционирования /Галиуллин К.Р., Обносова Н.А., Тухватуллина А.А., Шарипзянова Л.С. //Проблемы лексикологии и терминологии татарского языка. Вып.2. Казань, 1994. – С.127–134; *Бухараев Р.Г.* К концепции Машинного фонда Республики Татарстан /Бухараев Р.Г., Сафиуллина Ф.С., Галиуллин К.Р., Еникеев А.И., Сулейманов Д.Ш. //Татарский язык и новые информационные технологии. Вып.2. Казань: Изд-во Казан. ун-та, 1995. – С.20–35; *Володина Н.И.* Чувашская Республика /Многоязычие в России: региональные аспекты. М.: Межрегиональный центр библиотечного сотрудничества, 2008. – С. 22.; *Есипова А.В.* Создание машинного фонда шорского языка //Языки, духовная культура и история тюрков: традиции и современность (Труды международной конференции в 3-х томах). Т. 1. Казань, 1992. – С. 244–247.
6. *Пиотровский Р.Г., Щерба А.М., Гузев В.Г.* О создании машинного фонда тюркских языков // Советская тюркология, 1988, №2. – С.92–101.
7. Вторая Всесоюзная конференция по созданию МФ РЯ (Материалы конференции). 1988. – 230 с. Отв. ред. член-корр. АН СССР Ю.Н. Караулов; Материалы III

Всесоюзной конференции по созданию МФ РЯ. МГУ, 1990. – 148с. Под ред. С.Ф.Гилязова и Ю.Н.Караулова.
8. Адрес Машинного фонда русского языка http://www.cfrl.ru
9. *Сиразитдинов З.А., Надергулов И.У.* О создании Машинного фонда башкирского языка / Востоковедение в Башкортостане. История. Культура. Вып.3, Уфа, 1992; Сиразитдинов З.А. Компьютерная технология и языкознание / Компьютерные технологии в исторических и лингвистических исследованиях. Уфа, 1994.
10. Словарь говоров башкирского языка. Т.1, Уфа, 1967, 352 с. Словарь говоров башкирского языка. Т.2, Уфа, 1970, 326 с. Словарь говоров башкирского языка. Т.3. Уфа, 1987, 232 с. Диалектологический словарь абкширского языка. Уфа: Китап, 2002. с.
11. Диалектологический атлас башкирского языка. Уфа: Гилем, 2005, 234 с.
12. Грамматика современного башкирского литературного языка. Под ред. А.А.Юлдашева. М.: Наука, 1981.
13. *Сиразитдинов З.А.* Моделирование грамматики башкирского языка. Словизменительная система. Уфа: Гилем, 2006, 160 с.

ТҮЙІНДЕМЕ

Мақала тілдік ақпараттық қорды түркі тілдерінің салыстырмалы зерттеулеріне пайдалану мәселелеріне арналған. Автор башқұрт тілінің компьютерлік (машиналық) қор жиынтығын егжей-тегжейіне жеткізе қарастырады, олардың салыстырмалы және салғастырмалы зерттеулерде қолданылу мүмкіндіктерін анықтайды.
(Сиразитдинов З.А. Түркі тілдерінің салыстырмалы зерттеулерінің қайнар көзі ретіндегі башқұрт тілінің ақпараттық қоры)

SUMMARY

The article deals with the use of linguistic information resources in comparative studies of languages. The author considers in detail the resources of PC (computer) fund of the Bashkir language, identifies opportunities for the use of comparative and comparative studies. (**Sirazitdinov Z.A. Information resources Bashkir as a source for comparative studies of Turkic languages)**

ÖZET

Makalede dil bilgi fonunu Türk dillerinin karşılaştırmalı araştırmalarına kullanmak meseleleri söz konusudur. Yazar Başkurt dilinin bilgisayardaki fonunu detaylıca değerlendiriyor.
(Siraziddinov Z.A. Türk dillerinin mukayeseli araştırmalarının kaynağı olarak Başkurt dilinin bilgi fonu)