

БАШКИРСКИЙ ЯЗЫК И ЛИТЕРАТУРА В УСЛОВИЯХ ГЛОБАЛИЗАЦИИ И ПОЛИЭТНИЧЕСКОЙ СРЕДЫ: ОПЫТ И ПЕРСПЕКТИВЫ

ISBN 978-5-91608-181-7

С. 309-319

https://doi.org/10.31833/conf_ihll2019.060

УДК 8.11.512`374:004.651

З. А. Сиразитдинов

*Ордена «Знак Почета» Институт истории, языка и литературы УФИЦ РАН
проспект Октября, 71, Уфа, 450054, Россия
E-mail: sazin11@mail.ru*

ЭЛЕКТРОННЫЕ РЕСУРСЫ БАШКИРСКОЙ ЛЕКСИКОГРАФИИ

В статье рассматривается Машинный фонд башкирского языка (МФБЯ), разработанный в лаборатории лингвистики и информационных технологий Института истории, языка и литературы Уфимского федерального исследовательского центра Российской академии наук. МФБЯ является системой поиска филологической информации, разработанной для лингвистов, преподавателей, студентов и учащихся старших классов. Данная информационная система включает в себя специализированные базы данных, интерфейсы для выполнения запросов к базам и программные обеспечения для обработки этих запросов. Автор подробно описывает шесть разделов фонда, содержащих лексикографические базы данных и представляющих ценный ресурс для лексикографов в разработке новых словарей национального языка.

Наибольший интерес для составителей лексикографических трудов представляет раздел генеральной картотеки, на базе которой можно составить предварительный словник любого словаря. В этом разделе имеется информация о башкирской лексике с разных точек зрения: происхождения (исконная, заимствованная, языки заимствования), динамики (устаревшая, новая), стилистики (книжная, разговорная, просторечная), социальной принадлежности (диалектная, специальная, жаргонная) и др.

Лексикографический раздел МФБЯ дает информацию о слове исходя из существующих словарей башкирского языка. База данных этого раздела на сегодня включает 58 структурированных словарей с общим объемом в 382701 словарная статья. В разделе представлены академические и учебные словари: одноязычные и двуязычные, частотные, терминологические, синонимический, словари-справочники и другие. Все словари рассматриваемого раздела описаны в базах согласно принятым для групп структурам данных. В статье отмечается, что лексикографический раздел связан с разделом генеральной картотеки: для любого слова из картотеки можно определить список словарей в которых данное слово употреблено как реестровое слово или в словарных статьях (как определение или перевод) осуществить быстрый переход на эти словари.

В описании лексикографической базы диалектологического раздела отмечается возможность поиска информации по 6 информационным полям из 7 описанных для каждого диалектного слова.

В статье отмечается, что использование реляционных баз данных для обработки текстовых массивов имеет свои преимущества. Основное преимущество перевода текстов в табличный вид заключается в простоте и скорости создания утилит манипулирования данными. В частности, это

дает возможность реализации частотных словарей словоформ и основ для любых текстов или групп текстов, отобранных исследователем: авторские, стилевые, отдельных газет, фольклорных жанров и др. Построение частотных словарей в отечественной корпусной лингвистике реализовано впервые.

Автором отмечается, что лексикографические разделы МФБЯ активно и плодотворно используются лингвистами республики. В частности, это позволило сотрудникам отдела языкознания ИИЯЛ подготовить и издать десятитомный академический словарь башкирского языка за рекордно короткое время в 7 лет.

Ключевые слова: лингвистика, башкирский язык, лексикография, компьютерная лингвистика, лингвистические базы данных, компьютерная лексикография, корпусная лингвистика

Создание компьютерных словарей и размещение их в сети Интернет стало новым направлением современной лексикографии, которое, по крайней мере на нынешнем этапе, не направлено на полное замещение традиционных печатных лексикографических трудов. Такие словари нового типа не только расширяют сферу функционирования национального языка, но и открывают широкие перспективы для специалистов филологов в плане сравнительных и сопоставительных исследований лингвистических реалий языков в силу доступности лексикографических справочников.

В последнем случае есть и сдерживающие факторы в использовании информационных ресурсов. Таковыми являются: 1) объем филологической информации, 2) достоверность лингвистических данных.

Эти факторы касаются, в основном, информационных данных Википедии, малых коммерческих организаций и частных лиц. Но такие проблемы не возникают для лингвистических ресурсов, созданных на основе академических первоисточников в научных и научно-педагогических учреждениях, в крупных IT компаниях, тесно сотрудничающих с академическими институтами. Среди таких лексикографических разработок особо выделяются комплексные лингвистические ресурсы, включающие в себя множественные базы данных [СРЯ; Словари АБВУУ; Онлайн-словари FU-Lab; Словарный модуль]. Эти лингвистические ресурсы имеют большой и разнообразный филологический объем, их лексикографические базы дают возможность одновременного поиска в десятках и сотнях словарей одновременно.

Таким серьезным ресурсом в башкирской лингвистике является Машинный фонд башкирского языка (МФБЯ), содержащий разнообразную филологическую информацию и лексикографическую в частности. Данный фонд разработан в лаборатории лингвистики и информационных технологий Института истории, языка и литературы Уфимского федерального исследовательского центра Российской академии наук [МФБЯ].

Идея разработки МФБЯ, его основные положения были выдвинуты нами еще в начале 90-х годов прошлого столетия [Сиразитдинов и др. 1992]. Непосредственная работа по созданию фонда началась в конце 2003 г. с разработки концепции и структуры фонда как системы реляционных баз данных, работающей в локальной сети. В 2005 г. окончательно утвердилась концепция машинного фонда в виде открытой сетевой системы с доступом через Интернет, которая активно функционирует и пополняется новыми данными.

Сегодня МФБЯ является всеобъемлющей системой поиска филологической информации, разработанной для лингвистов, преподавателей, студентов и учащихся старших классов. Данная информационная система включает в себя специализированные базы данных, интерфейсы для выполнения запросов к базам и программные обеспечения для обработки этих запросов. Фонд разработан на основе СУБД Oracle и функционирует на специально выделенном сервере. На рис. 1 представлена структура информационной системы.

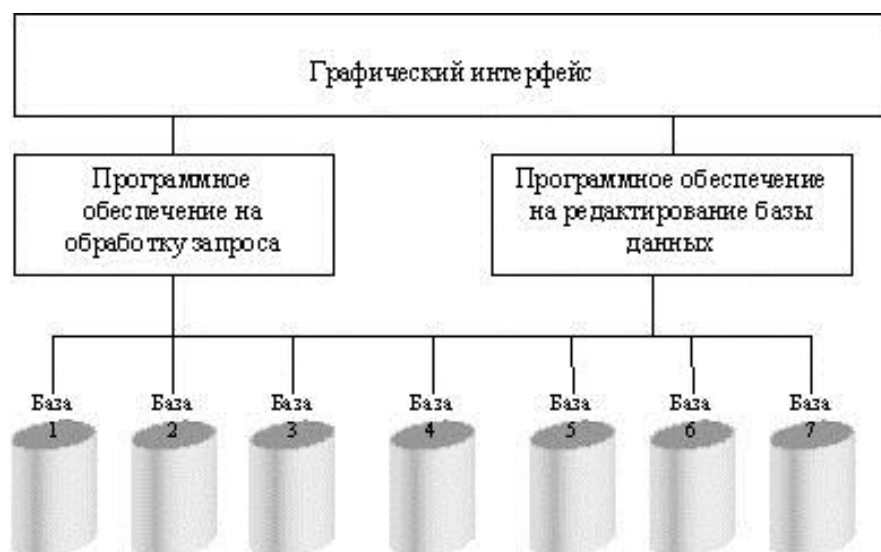


Рис. 1. Структура Машинного фонда башкирского языка.

На сегодня МФБЯ имеет 10 крупных разделов, которые представлены одним или несколькими раздельными базами данных. Из них 6 разделов содержат лексикографические данные, которые являются объектами нашего рассмотрения.

1. **Раздел генеральной картотеки** содержит основную информацию о лексической системе языка. В его базе в структурированном виде находятся более 100000 корнеслов и производных всех пластов языка.

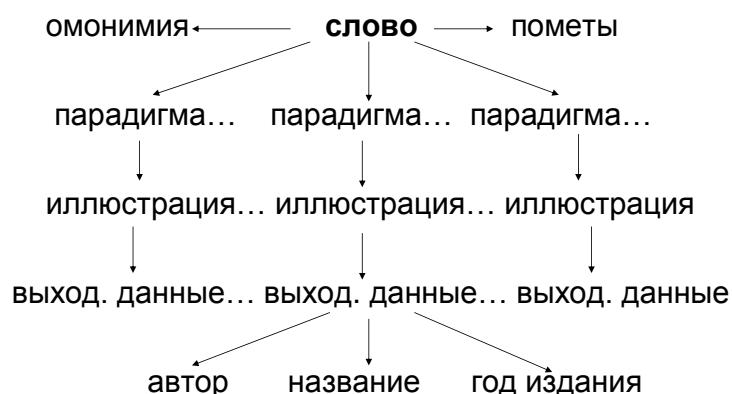


Рис. 2. Структура представления данных в генеральной картотеке.

Раздел дает информацию о башкирской лексике с разных точек зрения: происхождения (исконная, заимствованная, языки заимствования), динамики (устаревшая, новая), стилистики (книжная, разговорная, просторечная), социальной принадлежности (диалектная, специальная, жаргонная) и др. Разработаны более 50 помет, которые выбираются из списков при вводе новых или корректировке уже введенных слов в генеральную картотеку. Эта система помет призвана решить проблему функциональной и стилистической дифференциации лексики в башкирской лексикографии, которая до сих пор решается частично и несистематически. С освоением в общей лексикографии развернутой системы помет “отсутствие пометы в конкретных словарях становится уже значимым” [Юлдашев 1972: 405 с.].

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) из бумажной картотеки Института. На сегодня, когда МФБЯ содержит корпусные данные, информация о парадигмах является избыточной, но они оставлены в разделе как сбалансированные лингвистические данные, собранные и выверенные рядом поколений башкирских лексикографов в прошлом веке (30-90-е годы) и имеют историческое значение.

База данных этого раздела связана с базой лексикографического раздела, что позволяет определить в каких опубликованных словарях зафиксировано то или иное слово из генеральной картотеки.

2. **Лексикографический раздел** дает информацию о слове исходя из существующих словарей башкирского языка. База данных этого раздела на сегодня включает 58 структурированных словарей с общим объемом в 382701 словарная статья.

В данном разделе представлены академические и учебные словари: одноязычные и двуязычные, частотные, терминологические, синонимический, словари-справочники (названий населенных пунктов и улиц городов республики, названий водных и горных объектов Башкортостан) и другие. На рис. 2. показан вид интерфейса лексикографического раздела.

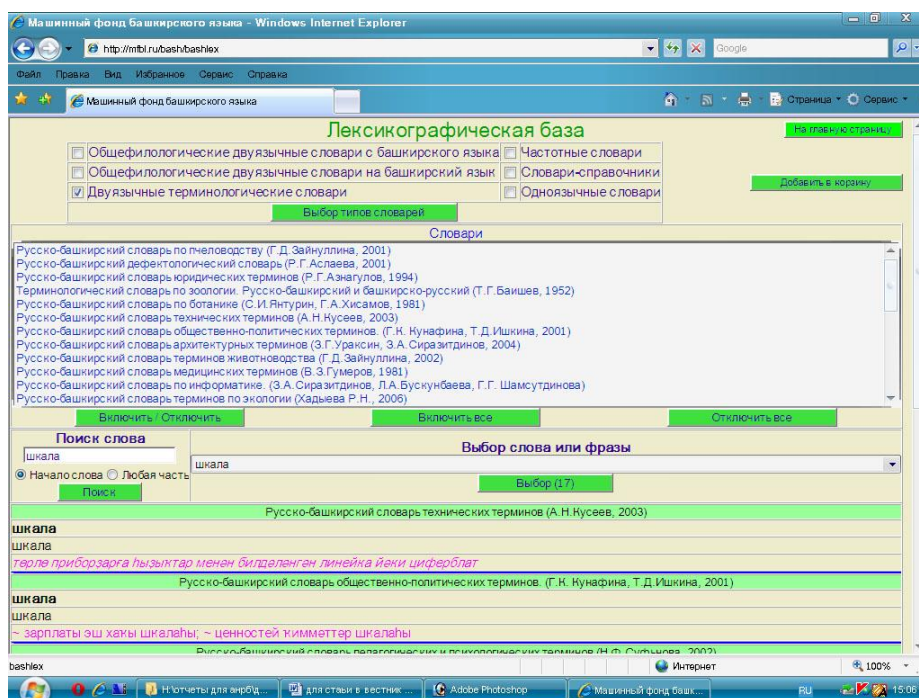


Рис. 2. Вид интерфейса лексикографического раздела.

Лексикографические базы данных рассматриваемого раздела МФБЯ построены на основе изданных в разные годы печатных словарей, с указанием авторов и года издания, за достоверность представленных данных лаборатория несет ответственность.

Создание электронного варианта отдельного словаря не представляет большой сложности. Построение же единой лексикографической системы из разных словарей, как отмечают разработчики, требует решения сложного вопроса интегрирования разноструктурных словарей [Очирова и др. 2012: 110].

Несмотря на существующие различия в структурах разных лексикографических разработок, их всех объединяет одно свойство: наличие опорного (реестрового) слова и наличие статьи толкования в виде ряда пояснительных полей. Это позволяет выделить обобщенную словарную структуру для групп однотипных лексикографических трудов,

внутри которой каждый словарь будет представлен наличием или отсутствием каких-либо информационных полей [Жубанов 2009: 213].

Все словари, которые вошли в лексикографический раздел, сгруппированы и описаны в виде структурных данных, имеющих от 4 до 11 информационных полей.

I. Частотные словари. Изданные 3 частотных словаря по функциональным стилям башкирского языка имеют единую структуру: 1) реестровое слово, 2) показатель части речи, 3) показатель омонимии, 4) частота употребления. Например: **Барлык** I И 39, **Барлык** II 1, **Һыр** II 3 [ЧСЛБЯП 2002].

II. Синонимический словарь. Данный словарь также имеет структурную схему состоящую из 4 информационных полей: 1) реестровое слово, 2) информация о синонимическом ряде, 3) толкование, 4) иллюстративный материал. Например: **Матур һылыу, күркәм, сибәр, нәфис** Күрер күзгә йәмле, күренеше, булмышы күнелгә ятышлы. Матур киң кулланылышлы: йән эйәләренә, кешегә, тәбиғәт күренештәренә карата әйтелә. *Әй матур кыз Зөһрә йән, мин һине касан күрәм?* (Ш. Бабич) [ССБЯ 2000]. Здесь после реестрового слова дается лексика (выделена курсивом), образующая со словарным словом синонимический ряд. После идет толкование реестрового слова и его отличительных свойств от остальных синонимичных слов. Четвертое поле дает пример употребления (выделено курсивом) реестрового слова в поэзии Ш.Бабича.

III. Орфографический словарь башкирского языка. Словарь имеет следующую структуру: 1) реестровое слово, 2) показатель омонимии, 3) показатель части речи, 4) дополнительная информация, 5) аффиксы словоизменения. Например: **Сап** I И *иск.* -тың, -ка, -бы(на), **Сап** II М, **Ағасак** И (балык) -тар; -ғы(на) [ОСБЯ 2002]. Дополнительная информация показывает к какому пласту лексики относится слово (устаревшее, диалектное, бранное) или уточняет и конкретизирует объект (вид, род, часть объекта).

IV. Двухязычные обычные терминологические словари. Обобщенную структуру опубликованных и вошедших в данный раздел русско-башкирских и башкирско-русских терминологических словарей (за исключением толковых терминологических словарей) можно описать в виде: 1) реестровое слово, 2) показатель многозначности, 3) специальная помета, 4) перевод, 5) дополнительная информация. Например: **Троеборье** *кон. спорт* есьярыш. **Толчок** 1. *тяж. атл.* күкрәктән сөйөү; 2. *науц.* *борьба* этеп эбәреү [РББРСФВС 2009]. **Капитель** *арх.* Капитель (*таш баганының, колоннаның югары өлөшө*) [РБСТТ 2003].

В некоторых словарях сложные термины, которые должны выделяться в виде отдельных словарных единиц, даются в словарных статьях как примеры словосочетаний. Такие термины выведены нами в отдельные реестровые слова.

V. Двухязычные толковые терминологические словари. Словари данного типа описываются структурой: 1) реестровое слово, 2) показатель многозначности, 3) специальная помета, 4) перевод, 5) толкование. Пример: **Убеждение** 1. ышаныу, инаныу (*шахестең ниндәйзер эш-максатты күңелгә ярашлы, мөһим, әһәмиәтле тип һанап аңлы ынтылышы. Күңелдә нык таралған караи*); 2. (*метод*) ышандырыу (*тәрбиәләнеүселәрзе укыу йә башка файзалы йәмғиәт эшмәкәрлегенә ылықтырыу йәки насар эш-кылыктан ситләштереү максатында уның аңына тәьсир итеүсе педагогик тәрбиә алымы*) [РБСППТ 2002].

Отдельными авторами указывается источник происхождения термина (латинское или греческое название термина). В лексикографическом разделе эта информация не учитывается, поскольку язык заимствования фиксируется в генеральной картотеке.

VI. Двухязычные словари-справочники (словари названий населенных пунктов и улиц городов, словари-справочники водных и горных объектов Республики Башкортостан). Эти 4 словаря-справочника описываются обобщенной структурой: 1) реестровое слово, 2) перевод, 3) рекомендуемое название, 4) указатель вида объекта, 5) территория расположения (протекания), 6) этимология. Рекомендуемое название встречается только для словаря названий населенных пунктов [РБССННПРБ 2001]. В словаре названий улиц городов отсутствует этимология [РБСНУГРБ 2009].

Например: **Чапаево** с. Сапай Сапай *Кидрячевский сельсовет Давлекановский район* от личн. им. Сапай [РБССННПРБ 2001].

Талкас оз. Талкас Баймакский район РБ, *возможно, от талкыш 'льномялка'* [РБСВОРБ 2005].

улица Новомостовая Яңы Түшәлгән юл урамы г.Уфа [РБСНУГРБ 2009].

VII. Толковый словарь современного башкирского литературного языка. Здесь рассматривается однотомный словарь, который представлен в лексикографическом разделе [ТСБЛЯ 2004]. Словарь имеет структуру: 1) реестровое слово, 2) показатель омонимии, 3) показатель части речи, 4) показатель многозначности, 5) толкование, 6) дополнительная информация, 7) иллюстративный материал в виде словосочетаний, 8) иллюстративный материал в виде фразеологизмов, пословиц и поговорок.

VIII. Двухязычные общие словари (башкирско-русский, русско-башкирский, башкирско-английский, англо-башкирский, башкирско-турецкий). Словари данного типа имеют самые сложные строения, которые обобщены нами к структуре: 1) реестровое слово, 2) показатель омонимии, 3) показатель многозначности, 4) дополнительная информация 1, 5) дополнительная информация 2, 6) дополнительная информация 3, 7) дополнительная информация 4, 8) перевод, 9) дополнительная информация 4, 10) иллюстративный материал в виде словосочетаний, предложений, 11) иллюстративный материал в виде фразеологизмов, пословиц и поговорок.

Так в башкирско-русском академическом словаре дополнительные информации до перевода выделяют диалектизмы, уменьшительные и ласкательные формы, прямые и переносные значения, стили употребления (разговорное просторечие), грамматические принадлежности и формы падежного управления, разъяснения употребления слов, формы управления при глаголах, косвенные падежи личных и указательных местоимений, залоговые и другие формы глаголов и др. [БРС 1996]. Дополнительная информация 4 после русского перевода дает различные дополнительные пояснения и уточнения значений и оттенков значения при переводе, падежные управления. Например:

Ете I 1. *числ. колич.* семь; **ете кеше** семь человек; **ете йыл** семь лет; *с аффиксом 3 л. ед. ч. поминки (у мусульман на седьмой день кончины) ...*

Ете II диал. 1) *см. үткер*; **ете бысак** острый нож; 2) *см. етеү II*.

Все словари рассматриваемого раздела описаны в базах согласно принятым для групп структурам данных. При выдаче на экран пустые поля не выводятся.

3. Лексикографическая база диалектологического раздела представляет информацию о диалектной лексике с указанием диалектов и говоров, в которых они распространены. База реализована на основе трехтомного словаря башкирских говоров [СБГ 1967; 1970; 1987]. Эти словари имеют следующие информационные поля: 1) реестровое диалектное слово, 2) показатель части речи, 3) указатель диалекта, 4) указатель говора, 5) нормативный литературный вариант, 6) русский перевод, 7) иллюстративный материал.

Например: Оланбармак исем, көнсығыш диалекты, мейәс һөйләше, һук бармак, указательный палец. **Суғыртмак** исем, көнсығыш диалект, суғырмак, кора (*нижней части ствола старой березы*). **Тумак** исем, төһьяк-көнбайыш диалект, ғәйнә һөйләше, тыумыш, рождение, *Тумактан калмаганга, үлмәктән калмабыз.*

В случае когда невозможен перевод на русский язык, дается толкование, например в этнографизмах: **Терзек** и. көнсығыш, урта урал, кыз яғынан кейәүгә әзерләнгән өс кейеме, одежда, приготовленная невестой для жениха *Терзеккә гел кижеләп бишбармак, кижеле ыштан, кижеле күлмәк әзерзәп һалып куйаһың инде һандығыңа.*

База данных построена соответственно вышерассмотренным информационным полям. Перевод или толкование диалектного слова помещается в одном поле "русское слово". Предоставляется поиск по первым 6 полям. На рис. 3.

Лексическая база данных

Поиск по полям

| | | |
|--------------------|----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| Башкирское слово | <input type="text"/> | <input checked="" type="radio"/> Любая часть <input type="radio"/> Слово точно <input type="radio"/> Начало слова <input type="radio"/> Конец слова |
| Литературная норма | <input type="text"/> | <input checked="" type="radio"/> Любая часть <input type="radio"/> Слово точно <input type="radio"/> Начало слова <input type="radio"/> Конец слова |
| Русское слово | золовка | <input checked="" type="radio"/> Любая часть <input type="radio"/> Слово точно <input type="radio"/> Начало слова <input type="radio"/> Конец слова |
| Часть речи | <input type="text"/> | |
| Диалект | <input type="text"/> | |
| Говор | <input type="text"/> | |

Поиск

Количество найденных слов: 45

| | |
|---------------------------|--------------------------------|
| Башкирское слово | БИКӨ |
| Часть речи | и. |
| Диалект | көнъяк |
| Говор | дым |
| Литературная норма | кәйенбиңә |
| Русское слово | золовка |
| Дополнительная информация | (старшая сестра мужа или жены) |
| Башкирское слово | БИКӨ |
| Часть речи | и. |
| Диалект | көнъяк |
| Говор | күртә |

Рис. 3. Пример выдачи из словарной базы диалектологического раздела.

4. Частотные словари корпусных проектов. Сейчас все крупные языки разработали свои национальные корпуса. Ведутся корпусные разработки и по языкам народов России: бурятскому [Бадмаева 2008], калмыцкому [НККЯ], лезгинскому [КЛЯ], осетинскому [КОЯ], татарскому [НКТЯ], тувинскому [Салчак 2012], шорскому [ЭЖШТ], хакасскому [Шеймович 2011] и др.

Учитывая актуальность корпусных разработок, сотрудниками лаборатории лингвистики и информационных технологий ведутся работы над созданием корпусов башкирского языка. На сегодня разработаны корпусные проекты прозаических, публицистических и фольклорных текстов башкирского языка с достаточным объемом текстов, которые представлены в МФБЯ самостоятельными разделами. Корпусы также реализованы на основе Oracle. Использование именно реляционных баз данных для обработки больших текстовых массивов имеет свои положительные стороны, к которым относится то, что реляционная база позволяет осмысленно разбивать тексты на элементарные составляющие, что приводит к появлению между всеми элементами связей. Создается возможность фильтрации и извлечения агрегированных характеристик текста при помощи одного или нескольких коротких (по сравнению с программой обработки текстовых файлов) запросов [Шевелев 2004: 222]. Основное преимущество перевода текстов в табличный вид заключается в простоте и скорости создания утилит манипулирования данными. В частности, это дало нам возможность реализации частотных словарей словоформ и основ для любых текстов или групп текстов, отобранных исследователем: авторские, стилевые, отдельных газет, фольклорных жанров и др. Построение частотных словарей в отечественной корпусной лингвистике реализовано впервые. выданный построенный корпус менеджером частотный словарь по рассказу А.Абушахманова “Юморески на сцене”.

Все вышерассмотренные лексикографические базы разделов МФБЯ активно используются лингвистами республики. В лексикографии сбору языкового материала всегда придается большое значение, так как чем больше иллюстративного материала, тем выше достоверность определения значения слова, тем шире сфера действия наблюдаемых закономерностей. По некоторым оценкам, для составления словаря в 30-40 тыс. слов необходима картотека в 10 миллионов цитат. Подготовка такого рода материала — рутинный и очень медленный процесс.

Сортировка

Основ в словаре 195

Сумма всех частот 388

Используются только словоформы со снятой омонимией

| | | | | | | | | | |
|-------------|--------------|-------------|------------|-------------------|----------------|------------|------------|------------|-----------|
| агай 9 | сәхнә 9 | ти 9 | бул 8 | спектакль 8 | уйна 8 | баш 7 | кәзә 7 | менән 7 | өр 7 |
| тауыш 7 | бер 6 | мин 6 | сык 6 | тип 6 | был 5 | ла 5 | режиссер 5 | һәм 5 | артист 4 |
| башла 4 | көл 4 | роль 4 | ул 4 | һейәрғолов 4 | шул 4 | ә 4 | әйт 4 | ал 3 | икән 3 |
| ит 3 | кит 3 | кайт 3 | сығар 3 | театр 3 | төрлө 3 | тәкә 3 | һин 3 | шулай 3 | эшлә 3 |
| юк 3 | азак 2 | арт 2 | барыш 2 | башкар 2 | бөрийән 2 | бөтөр 2 | гена 2 | ғәзим 2 | йөрөт 2 |
| Һәшә 2 | күк 2 | күкрә 2 | кош-корт 2 | местком 2 | мысҡыл 2 | мысҡылла 2 | нишлә 2 | тамашасы 2 | тик 2 |
| тиклем 2 | тор 2 | фәкәт 2 | хужа 2 | хушлаш 2 | хәл 2 | эргә 2 | эт 2 | яз 2 | якшы 2 |
| аға 1 | акыр 1 | аркан 1 | артистка 1 | асыулан 1 | бакыр 1 | баһтыр 1 | башкаса 1 | башкорт 1 | бөз 1 |
| бел 1 | бик 1 | бир 1 | бирел 1 | бирле 1 | булдыр 1 | буһат 1 | бәйлән 1 | вакыт 1 | ғиз 1 |
| ғәзәт 1 | да 1 | дәмбәслә 1 | за 1 | зә 1 | ер 1 | еткер 1 | еткәлә 1 | етәксә 1 | заман 1 |
| зыяевич 1 | изән 1 | ике 1 | ила 1 | ирек 1 | иртәғәһән 1 | иҫәпән 1 | йый 1 | йыйыл 1 | йыл 1 |
| кейен 1 | кем 1 | кеүек 1 | кил 1 | күрен 1 | күтәр 1 | кәрәк 1 | кабул 1 | калай 1 | кара 1 |
| кас 1 | кат 1 | катын-кыз 1 | кола 1 | которт 1 | куй 1 | куш 1 | күш 1 | кыз 1 | кәнәғәт 1 |
| лә 1 | магнитофон 1 | макта 1 | мал 1 | мәһрә 1 | музыка 1 | мылтык 1 | мәжбүр 1 | мәл 1 | мәскәү 1 |
| мәсәлә 1 | мәсәлән 1 | на 1 | нимә 1 | ниңдәй 1 | ниһайәт 1 | овчарка 1 | озат 1 | озон 1 | ошо 1 |
| радио 1 | район 1 | сак 1 | саҡыр 1 | санкт-петербург 1 | сәнки 1 | такта 1 | тамамлан 1 | тар 1 | тартыл 1 |
| таҫма 1 | тиз 1 | тикшер 1 | током 1 | төрттөр 1 | тыуған 1 | тыш 1 | тә 1 | ук 1 | уқы 1 |
| урман-тау 1 | үз 1 | үс 1 | хәзәр 1 | хәйзәрөв 1 | һайра 1 | һәйләш 1 | һөт 1 | һуктыр 1 | һуңынан 1 |
| һүз 1 | һын 1 | шаршау 1 | шелтә 1 | ызалат 1 | әлек-әлектән 1 | әләк 1 | эш 1 | әгәр 1 | әйзә 1 |
| әлбитгә 1 | әлеге 1 | як 1 | яп 1 | ярҙамсы 1 | | | | | |

Рис. 4. Пример выдачи частотного словаря основ по выбранному тексту.

Когда же исходный продукт (текст) и конечный (картотека) находятся в машине, то скорость убыстряется в сотни раз. Поэтому лексикографические разделы активно использовались сотрудниками отдела языкознания ИИЯЛ в создании десяти томного академического словаря. Это позволило подготовить и издать данный словарь за рекордно короткое время в 7 лет (2011-2018 гг.).

Лексикографические данные МФБЯ сегодня представляют интегрированную автоматизированную картотеку для построения новых словарей. И как отмечают лексикографы, любая картотека должна быть полной во всех ее частях, должна адекватно отражать лексико-фразеологический состав языка в рассматриваемый временной период [Рогожникова и др. 1984: 20]. Это требует от сотрудников, занятых сопровождением лингвистических информационных систем, постоянной работы над пополнением состава баз данных, выявлением опечаток и ошибок в самих данных, которые по причине большого объема информации все же встречаются. Сегодня сотрудники лаборатории лингвистики и информационных технологий ИИЯЛ, наряду с решением новых задач, также выполняют и выше обозначенные задачи по поддержке МФБЯ.

Литература

Бадмаева и др. 2008 — Бадмаева Л.Д., Бадагаров Ж.Б., Цыдыпов Б.З. Общие проблемы формирования корпуса бурятского языка //Труды международной конференции «Корпусная лингвистика – 2008» 6–10 октября 2008 г., Санкт-Петербург. Санкт-Петербург, 2008. С. 24-30

БРС 1996 — Башкирско-русский словарь/под ред. З.Г.Ураксина. М.:Дигора, 1996. 884 с.

Гаджихмедов 2013 — Гаджихмедов Н.Э. На пути создания диалектного корпуса кумыкского языка// Актуальные проблемы диалектологии языков народов России: Материалы XIII международной конференции. Уфа, 2013. С. 177-179.

Жубанов 2009 — Жубанов А.К. База данных “Тіл — қазына” казахского слова и ее теоретические основы. Алматы: Изд. “Арыс”, 2009. 304 с.

КЛЯ — Корпус лезгинского языка [электронный ресурс] URL:<http://www.dag-languages.org/LezgianCorpus/search/> (дата обращения: 17.03.2019).

КОЯ — Корпус осетинского языка [электронный ресурс] URL:http://www.ossetic-studies.org/iron-corpus/search/index.php?interface_language=ru. (дата обращения: 17.03.2019).

- МФБЯ — Машинный фонд башкирского языка [электронный ресурс] URL:<http://mfb12.ru> (дата обращения: 17.02.2019).
- НККЯ — Национальный корпус калмыцкого языка [электронный ресурс] URL:<http://kalmcorp.org.ru/> (дата обращения: 17.02.2019).
- НКТЯ — Национальный корпус татарского языка [электронный ресурс] URL:<http://corpus.antat.ru> (дата обращения: 17.02.2019).
- Онлайн-словари FU-Lab — Онлайн-словари FU-Lab (финно-угорских языков) [электронный ресурс] URL:<http://dict.fu-lab.ru> (дата обращения: 17.01.2019).
- ОСБЯ 2002 — Орфографический словарь башкирского языка/ Н.Ф. Суфьянова, З.К. Ишмухаметов. Уфа: НИ "Башкирская энциклопедия", 2002. 352 с.
- Очирова и др. — Очирова Н.Ч., Куканова В.В. К вопросу о разработке словарного модуля в национальном корпусе калмыцкого языка// Научное наследие проф. А.Ш. Кичикова и актуальные проблемы современной калмыцкой филологии и культуры. Элиста: Изд-во Калмыцкого гос.ун-та, 2012. С. 108–110.
- РББРСФВС 2009 — Русско-башкирский, башкирско-русский словарь по физическому воспитанию и спорту/ Г.Д. Зайнуллина. Уфа: Китап, 2009. 160 с.
- РБСВОРБ 2005 — Русско-башкирский словарь водных объектов Республики Башкортостан /Ф.Г.Хисамитдинова, Р.З.Шакуров, З.А.Сиразитдинов, Ю.Х.Юлдашбаев. Уфа: Китап, 2005. 256 с.
- РБСНУГРБ 2009 — Русско-башкирский словарь-справочник названий улиц городов Республики Башкортостан/ Ф.Г.Хисамитдинова, З.А.Сиразитдинов, Н.Ф.Суфьянова. Уфа: ГУП РБ «Уфимский полиграфкомбинат», 2009. 312 с.
- РБССННПРБ 2001 — Русско-башкирский словарь-справочник названий населенных пунктов Республики Башкортостан/ Ф.Г. Хисамитдинова, З.А. Сиразитдинов. Уфа: Китап, 2001. 320 с.
- РБСППТ 2002 — Русско-башкирский словарь педагогических и психологических терминов/ Н.Ф. Суфьянова. Уфа: Китап, 2002. 112 с.
- РБСТТ 2003 — Русско-башкирский словарь технических терминов/ А.Н. Кусеев. Уфа: Башкирское издательство «Китап», 2003. 384 с.
- Рогожникова 1984 — Рогожникова Р.П., Чернышева Л.В. Организация словарной картотеки на базе автоматизированной системы//Теория и практика современной лексикографии. Л.:Наука, 1984. С. 20-27.
- Салчак А. Я. 2012 — Электронный корпус текстов тувинского языка // Новые исследования Тувы. 2012, № 3. (Электронный журнал). URL:http://www.new-tuva.info/journal/issue_15/5231-salchak.html (дата обращения: 17.03.2019).
- СБГ — Словарь башкирских говоров. В 3-х томах. Т.1, Уфа, 1967. 300 с. Т.2, Уфа, 1970. 327 с. Т.3, Уфа, 1987. 231 с.
- Сиразитдинов и др. 1992 — Сиразитдинов З.А., Надергулов И.У. О создании Машинного фонда башкирского языка // Востоковедение в Башкортостане. История. Культура. Вып.3, Уфа, 1992. С. 104-106.
- СРЯ — Словари русского языка [электронный ресурс] URL:<http://gramota.ru> (дата обращения: 17.02.2019).
- Словари АБВУ — Словари АБВУ LINGVO [электронный ресурс] URL:<https://www.lingvolive.com/ru-ru> (дата обращения: 17.02.2019).
- Словарный модуль Национального корпуса калмыцкого язык. [электронный ресурс] <http://bilib.ru/dictionary/search> (дата обращения: 17.02.2019).
- ТССБЛЯ 2004— Толковый словарь современного башкирского литературного языка/под. ред. З.Г. Ураксина. Уфа, 2004. 528 с.
- ЧСБЯП 2002 — Частотный словарь башкирского языка. Т.1 (проза)/ З.А.Сиразитдинов. Уфа: Гилем, 2002. 413 с.
- Шевелев 2004 — Шевелев О.Г. Представление набора текстов в реляционной базе данных для целей лингвистического анализа//Вестник Томского государственного университета. 2004, № 284. С.222-226.
- Шеймович 2011 — Шеймович А. В. Морфологическая разметка корпуса хакасского языка //Российская тюркология. 2011, № 2(5). С. 48–61.
- ЭКШТ — Электронный корпус шорских текстов [электронный ресурс] URL:<http://shoriya.ngpi.rdtc.ru> (дата обращения: 17.06.2013).
- Юлдашев А.А. Принципы составления тюркско-русских словарей. М.: Наука, 1972. 416 с.

Z. A. Sirazitdinov

*Order of the Badge of Honour Institute of History,
Language and Literature UFRC Russian Academy of Sciences
October Avenue, 71, Ufa, 450054, Russia
E-mail: sazin11@mail.ru*

ELECTRONIC RESOURCES OF BASHKIR LEXICOGRAPHY

The article discusses the Machine Fund of the Bashkir Language (MFBL), developed in the laboratory of linguistics and information technology of the Institute of History, Language and Literature of the Ufa Federal Research Center of the Russian Academy of Sciences. MFBL is a search system for philological information developed for linguists, teachers, students and high school students. This information system includes specialized databases, interfaces for querying databases and software for processing these requests. The author describes in detail the six sections of the fund, which contain lexicographical databases and represent a valuable resource for lexicographers in developing new dictionaries of the national language.

Of greatest interest to compilers of lexicographic works is the section of the general card file, on the basis of which a preliminary vocabulary of any dictionary can be compiled. In this section there is information about the Bashkir vocabulary from different points of view: origin (primordial, borrowed, borrowing languages), dynamics (obsolete, new), stylistics (book, conversational, colloquial), social affiliation (dialectal, special, slang) and others .

The lexicographic section of the MFBL provides information about a word based on existing dictionaries of the Bashkir language. The database of this section today includes 58 structured dictionaries with a total of 382701 entries. The section presents academic and educational dictionaries: monolingual and bilingual, frequency, terminological, synonymic, reference dictionaries and others. All the dictionaries of the section in question are described in the databases according to the data structures adopted for the groups. The article notes that the lexicographic section is associated with the section of the master card file: for any word from the card file, you can define a list of dictionaries in which the word is used as a register word or in dictionary entries (as a definition or translation) to quickly switch to these dictionaries.

In the description of the lexicographic base of the dialectological section, the ability to search for information on 6 information fields out of 7 described for each dialect word is noted.

The article notes that the use of relational databases for processing text arrays has its advantages. The main advantage of translating texts into a tabular view is the simplicity and speed of creation of data manipulation utilities. In particular, it enables the implementation of frequency dictionaries of word forms and bases for any texts or groups of texts selected by the researcher: author, style, individual newspapers, folklore genres, etc. The construction of frequency dictionaries in Russian corpus linguistics is implemented for the first time.

The author notes that the lexicographic sections of the MFBL are actively and fruitfully used by the linguists of the republic. In particular, it allowed the staff of the Department of Linguistics of the Institute of Language Studies to prepare and publish a ten-volume academic dictionary of the Bashkir language in a record short time of 7 years.

Keywords: linguistics, Bashkir language, lexicography, computational linguistics, linguistic database, computational lexicography, corpus linguistics.

References

- Badmayeva and others. 2008 - Badmayeva L. D., Badagarov Zh. B., Tsydypov B. Z. General problems of Buryat language corpus formation // Proceedings of the international conference "Corpus linguistics-2008" October 6-10, 2008, St. Petersburg. St. Petersburg, 2008. Pp. 24-30
- BRS 1996-Bashkir-Russian dictionary / edited by Z. G. Uraksin. Moscow: Digora, 1996. 884 p.
- Gadzhiahmedov 2013 — Gadzhiahmedov N. E. Towards creating a dialect corpus of Kumyk language // Actual problems of dialectology of languages of the peoples of Russia: Materials of XIII international conference. Ufa, 2013. 177-179 p.

- Zhubanov 2009 — Zhubanov A. K. Database "Til-Kazyra" of the Kazakh word and its theoretical foundations. Almaty: Ed. "Arys", 2009. 304 p.
- KLYA — Corpus of the Lezgin language [electronic resource] URL:[http:// www. dag-languages. org/LezgianCorpus/search/](http://www.dag-languages.org/LezgianCorpus/search/) (date of access 17.03.2019).
- KOYA — Corpus of Ossetian language [electronic resource] URL:[http:// www. ossetic-studies. org/iron-corpus/search/index.php? interface_language=ru](http://www.ossetic-studies.org/iron-corpus/search/index.php?interface_language=ru). (accessed 17.03.2019).
- MFBL — Machine Fund of the Bashkir language [electronic resource] URL: [http:// mfbl2. ru](http://mfbl2.ru) (date of access 17.02.2019).
- NKKYA — National corpus of Kalmyk language [electronic resource] URL: [http:// kalmcorporu.ru/](http://kalmcorporu.ru/) (date of access 17.02.2019).
- NKTYA — National corpus of the Tatar language [electronic resource] URL:[http:// corpus. antat. ru](http://corpus.antat.ru) (date of access 17.02.2019).
- Online dictionaries FU-Lab — Online dictionaries FU-Lab (Finno-Ugric languages) [electronic resource] URL:[http:// dict. fu-lab. ru](http://dict.fu-lab.ru) (date of access 17.01.2019).
- OSBYA 2002 — Spelling dictionary of the Bashkir language / N. F. Sufyanova, Z. K. Ishmukhametov. Ufa: NI "Bashkir encyclopedia", 2002. 352 p.
- Ochirova and others. — Ochirova N. CH., Kukanova V. V. On the development of the vocabulary module in the national corpus of the Kalmyk language // Scientific heritage of Prof. A. Sh. Kichikov and actual problems of modern Kalmyk Philology and culture. Elista: Publishing house of Kalmyk state University, 2012. pp 108-110.
- Russian-Bashkir, Bashkir-Russian dictionary of physical education and sport / G. D. Zainullina. Ufa: Kitap, 2009. 160 p.
- Russian-Bashkir dictionary of water objects of the Republic of Bashkortostan / F. G. Khisamitdinova, R. Z. Shakurov, Z. A. Sirazetdinov, Yu. Kh. Yuldashbaev. Ufa: Kitap, 2005. 256 p.
- Russian-Bashkir dictionary-directory of street names of cities of the Republic of Bashkortostan / F. G. hisamitdinova, Z. A. Sirazitdinov,, N. F. Sufyanova. Ufa: "Ufa polygraph plant", 2009. 312 PP.
- Russian-Bashkir dictionary-directory of names of settlements of the Republic of Bashkortostan / F. G. Khisamitdinova, Z. A. Sirazitdinov. Ufa: Kitap, 2001. 320 p
- Russian-Bashkir dictionary of pedagogical and psychological terms / N. F. Sufyanova. Ufa: Kitap, 2002. 112 p.
- Russian-Bashkir dictionary of technical terms / A. N. Kuseev. Ufa: Bashkir publishing house "Kitap", 2003. 384 p.
- Rogozhnikova 1984 – Rogozhnikova R. P., Chernysheva L. V. Organization of dictionary card index on the basis of automated system // Theory and practice of modern lexicography. L.: Nauka, 1984. p. 20-27.
- Salchak A. Ya. 2012 — Electronic corpus of texts of the Tuvan language // New studies of Tuva. 2012, No. 3. (Electronic journal.) URL:http://www.new-tuva.info/journal/issue_15/5231-salchak.html (date of access: 17.03.2019).
- Dictionary of Bashkir dialects. In 3 volumes. Vol. 1, Ufa, 1967. 300 S. T. 2, Ufa, 1970. 327 S. T. 3, Ufa, 1987. 231 p.
- Sirazitdinov and others. 1992 — Sirazitdinov Z. A., Nadergulov I. U. On the creation of the Machine Fund of the Bashkir language // Oriental Studies in Bashkortostan. History. Culture. Vol.3, Ufa, 1992. 104-106 p.
- Dictionaries of the Russian language [electronic resource] URL: [http:// gramota. ru](http://gramota.ru) (date of access 17.02.2019).
- Dictionaries ABBYY — Dictionaries ABBYY LINGVO [electronic resource] URL:[https:// www. lingvolive. com/ru-ru](https://www.lingvolive.com/ru-ru) (date of access 17.02.2019).
- Dictionary module of the national corpus of Kalmyk language. [electronic resource] <http://biliq.ru/dictionary/search> (date of access: 17.02.2019).
- Explanatory dictionary of the modern Bashkir literary language / pod. ed. Ufa, 2004. 528 p.
- Frequency dictionary of the Bashkir language. Vol. 1 (prose) / Z. A. Sirazitdinov. Ufa: Gilem, 2002. 413 p.
- Shevelev 2004 — Shevelev O. G. Presentation of a set of texts in a relational database for the purposes of linguistic analysis. Vestnik Tomsk state University. 2004, No. 284. p. 222-226.
- Sheimovich 2011 — Sheimovich A. V. Morphological marking of the Khakass language corpus // Russian Turkology. 2011, No. 2 (5). 48-61 p.
- Electronic corpus of Shor texts [electronic resource] URL: [http:// shoriya. ngpi.rdtc. ru](http://shoriya.ngpi.rdtc.ru) (date of access 17.06.2013).
- Yuldashev A. A. Basis of preparation of the Turkic-Russian dictionaries. Moscow: Nauka, 1972. 416 p.