EUROASIA JOURNAL OF SOCIAL SCIENCES & HUMANITIES
INTERNETIONAL INDEXED & REFEREED
ISSN 2651-5261

РАЗРАБОТКА КОРПУСА ТЕКСТОВ АФОРИСТИЧЕСКИХ ЖАНРОВ БАШКИРСКОГО ЯЗЫКА[1]

Л. А. БУСКУНБАЕВА (L. BUSKUNBAEVA)
З. А. СИРАЗИТДИНОВ (Z. SIRAZITDINOV)
А. Ш. ИШМУХАМЕТОВА (A. ISHMUKHAMETOVA)
Г. Г. ШАМСУТДИНОВА (G. SHAMSUTDINOVA)

Аннотация
В статье рассматриваются принципы разработки корпуса текстов афористических жанров башкирского фольклора. Включенные в базу данных тексты малых жанров фольклора станут источником для исследования языка фольклора. Авторы подробно останавливаются на проблеме аннотирования данных текстов. Рассматриваются возможности использования корпуса в научных трудах, учебном процессе.

Ключевые слова: Башкирский Язык, Тюркские Языки, Корпусная Лингвистика, База Данных, Фольклор, Афористические Жанры, Система Разметок.

Abstract
The article discusses the principles of the design of corpora of texts aphoristic genres in the Bashkir folklore. Small genres of folklore included in the database will allow the reader to explore the language of folklore. The authors dwell in detail on the problem of annotating data texts. Also the article discusses the possibility of using the developed case in the scientific works and educational process.

1. Введение

Фольклорные материалы, отражающие быт и мировоззрение народа, содержащие архаические элементы, являются ценным источником как для теоретических исследований истории развития языка, определения языковой картина мира, так и для практической лексикографии. Поэтому в отечественной корпусной лингвистике интерес к текстам народного творчества с каждым днем возрастает: активно ведутся работы по созданию корпусов фольклорных текстов русского [Николаев, 2015], калмыцкого [Куканова, 1912], нганасанского [Корпус нганасанских фольклорных текстов], тувинского [Салчак, 2012] языков. Институтом этнологии и антропологии разрабатываются корпусы фольклора ряда языков Сибири (эвенкийского, шорского, ненецкого, телеутского) [Корпусы ИЭА РАН]. Объектами этих корпусов выступают эпические, сказочные, библейские и мифологические тексты. Афористический жанр до сегодняшнего дня ни в одном из языков народов России не является объектом построения корпуса, корпусного исследования. Данная проблема впервые поднимается башкирскими языковедами.

Афористические жанры башкирского народного творчества – небольшие по объему фольклорные произведения, которые включают в свой состав пословицы (мәҡәлдәр), поговорки (әйтемдәр), загадки (йомаҡтар), приметы (һынамыштар), запреты (тыйыуҙар), предсказания (юрауҙар) и т.д. В лаконичных и емких суждениях находят отражение жизненные наблюдения и правила житейской мудрости башкирского народа.

[1] “Исследование выполнено при финансовой поддержке РФФИ и Правительства Республики Башкортостан в рамках научного проекта № 17-14-02010 а/р”.

Сотрудниками отдела фольклористики и искусства Института истории, языка и литературы Уфимского научного центра Российской академии наук (далее – ИИЯЛ УНЦ РАН) во время многочисленных экспедиций по районам Республики Башкортостан, соседним областям и республикам, где компактно проживает башкирское население, был собран богатый материал по афористическим жанрам башкирского фольклора. Данный материал представлен в виде отдельных томов “Башкирского народного творчества” [Башҡорт халыҡ ижады, 1995; Башҡорт халыҡ ижады, 2007; Башҡорт халыҡ ижады, 2006], словарей [Башҡортса-русса фразеологик һүҙлек, 1973; Духовное наследие.., 2008; Башкирско-англо-русский словарь адекватных пословиц и поговорок, 2002] и монографий [Надршина Ф.А., 2008; Нәҙершина, 1983].

Объемы собираемого материала по фольклору растут, и в основном они представлены на бумажных носителях. Создание электронной базы данных текстов афористических жанров, интегрированной в корпус фольклора башкирского языка, позволяет зафиксировать их в единой базе и пополнять по мере фиксации. Данный проект предоставляет возможность решать задачи сохранения культурного наследия башкирского народа с помощью новых технологий и способствовать широкому распространению материалов башкирского фольклора в ознакомительных и научных целях. Наличие тегирования корпуса позволяет проводить статистический и автоматизированный анализы текста, что, в свою очередь, дает возможность исследовать конкретное слово в синхронном и диахронном аспектах, получать данные о частоте лексем и грамматических категорий, о сочетаемости лексических единиц.

2. Архитектура корпуса

Данный корпус разрабатывается в русле общей концептуальной модели корпуса башкирского языка (КБЯ), которая включает на сегодняшний день в себя подкорпусы прозаических, публицистических (газетных и журнальных) и фольклорных (эпических и сказочных) текстов [Бускунбаева 2011, 45–51; Бускунбаева 2012, 139–141; Бускунбаева 2012, 54–58; Бускунбаева 2013, 135–140; Сиразитдинов 2014, 86–89; Сиразитдинов, 2015, 658–664; Сиразитдинов 2013; Сиразитдинов 2011, 269–274].

Для функционирования КБЯ в сети Интернет лабораторией лингвистики и информационных технологий ИИЯЛ УНЦ РАН разработана интегрированная система, позволяющая создавать корпусы, осуществлять широкий круг поисковых задач и администрирования баз данных [Сиразитдинов, Полянин 2014; Sirazitdiniv, 2014]. Она разработана на основе системы управления базами данных ORACLE.

Интегрированная система состоит из двух блоков: пользовательский и администраторский.

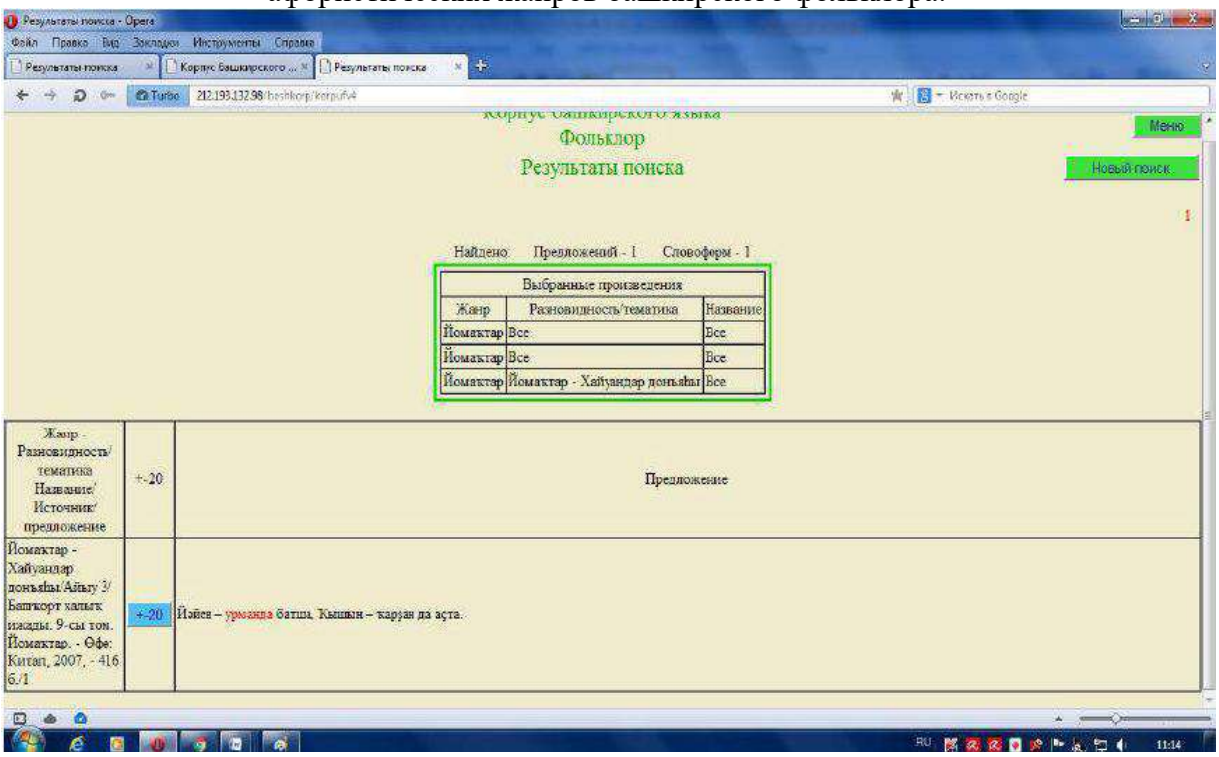
I. Пользовательский блок включает следующие программные средства:

1. Средства определения объема корпуса, выделения пользовательского подкорпуса.
2. Поисковые средства. Программы поиска позволяют производить гибкий поиск по многим лингвистическим параметрам: словоформе, лемме, сочетанию лемм, сочетанию словоформ (от двух и выше по желанию пользователя), грамматическим категориям, сочетанию грамматических категорий (до пяти), семантике, сочетанию семантик. Данный корпус также позволяет строить частотные словари основ и словоформ по любому тексту, по совокупности текстов, относящихся к определенному типу.

II. Блок администратора (с правом входа для сотрудников лаборатории) включает следующие программные средства:

1. Программные средства ввода и автоматической разметки текстов. Данные средства производят морфологические и семантические разметки введенных новых текстов.
2. Средства редактирования. Предусмотрены возможности редактирования основного словаря, списков словоизменительных категорий, моделей словоизменения и правка самих текстов.
3. Средства ручного снятия грамматических и лексических неоднозначностей. Сотрудники лаборатории могут просматривать текст по предложениям и устранять омонимичные явления, которые не разрешаются самой системой.
4. Программы статистического учета посещаемости корпуса текстов.
5. Программа экспорта любого размеченного текста из базы данных ORACLE в формате xml для обмена данными с другими корпусными проектами.

Рис. 1. Интерфейс корпуса фольклорных текстов с привлечением текстов афористических жанров башкирского фольклора.

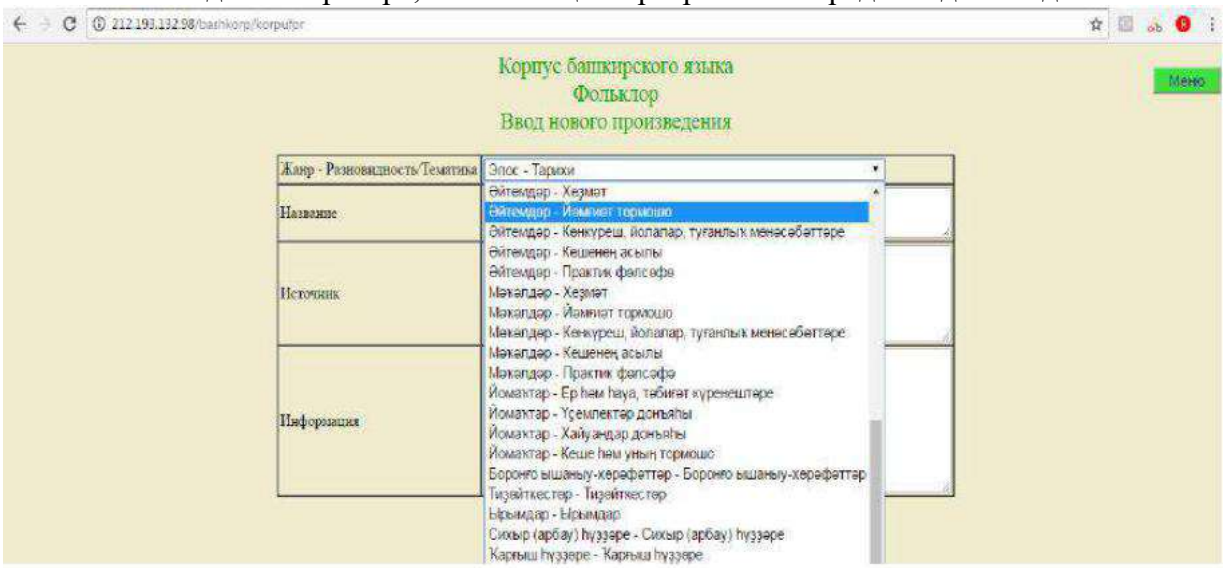


3. Система разметок текстов афористических жанров башкирского фольклора

На сегодняшний день разразработаны метатекстовая и лингвистическая (морфологическая) разметки текстов афористических жанров башкирского фольклора.

Метатекстовая разметка текстов афористических жанров (информация о тексте):
– **жанр** (пословица „мәкәл“, поговорка „әйтем“, загадка „йомаҡ“, примета „һынамыш“, запрет „тыйыу“, предсказание „юрау“);
– **тематика** (например, для загадок представлены следующие виды: земля и небо, явления природы „ер һәм күк, тәбиғәт күренештәре“, мир растений „үҫемлектәр донъяһы“, мир животных „хайуандар донъяһы“, человек и его жизнедеятельность „кеше һәм уның тормошо“);
– **название** (если имеется, например, в загадках в качестве названия дается его ответ);
– **источник** (название источника, год издания);
– **объем текста** (количество предложений, словоформ).

Рис.2. Блок администратора, включающий программные средства для ввода текстов.



Система морфологической разметки в данном корпусе ориентирована на представление всех регулярных словоизменительных грамматических форм.

Морфологическая информация башкирской словоформы в корпусе включает:
– исходную форму слова (лемму);
– частеречную характеристику;
– совокупность морфологических признаков по типу агглютинативных аффиксов словоизменения, которые подразделяются на именные и глагольные формы.

Выделяются 12 частей речи: имена существительные, числительные, прилагательные, наречия, глаголы, местоимения, подражательные слова, междометия, модальные слова, союзы, частицы, послелоги.

Принятые обозначения категорий частей речи:
N (noun) – существительное: *бала, шатлык;*
V (verb) – глагол: *барыу, тырышыу;*
NUM (numeral) – числительное: *ике, алтмышар;*
ADV (adverb) – наречие: *яй, арттан;*
ADJ (adjective) – прилагательное: *бәләкәй, ҡыҙыл;*
PRON (pronoun) – местоимение: *беҙ, ҡасан;*
POST (postposition) – послелог: *менән, өсөн;*
CONJ (conjunction) – союз: *әммә, йәиһә;*
PART (particle) – частица: *генә, әле;*
INTJ (interjection) – междометие: *ай, тфу;*
MOD (modal word) – модальное слово: *эйе, юҡ, түгел, әлбиттә;*
IMIT (imitative word) – подражательное слово: *сылтыр, шалтыр.*
Именные морфологические признаки включают показатели 15 категорий:
• категория числа (единственное число Sg `singular` и множественное число Pl „plural“),
• категория падежа (основной падеж Nom `nominative`, родительный Gen `genitive`, дательный Dat `dative`, винительный Acc `accusative`, исходный Abl `ablative`, местный падеж Loc `locative`),
• категория принадлежности (Poss `possessive`:*-м/-ң/ -һы/-быҙ/-гыҙ*),
• категория сказуемости (Pred `predicativity`:*-мын/-һыңˈ-быҙ/-һығыҙ*),

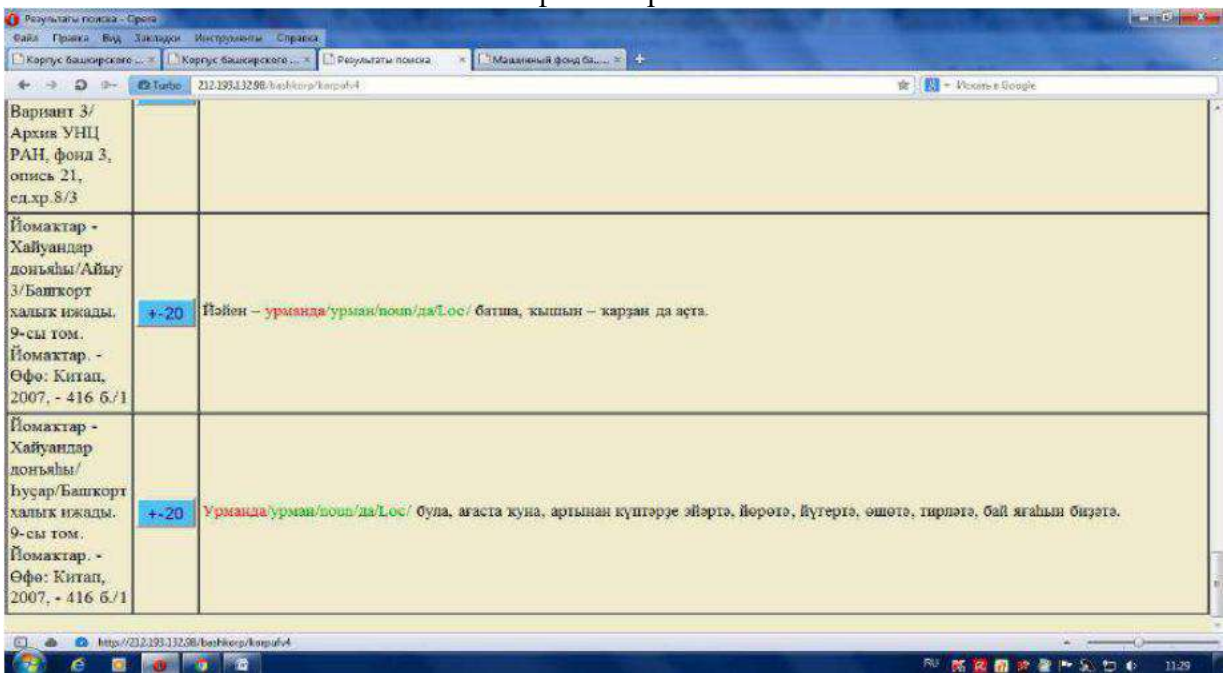
• категория вопросительности (Q `question`: *-мы/-ме*),
• категория неопределенности (Indf `indefinite`:*-дыр/-дер*),
• категория усиления (Int `intensifying`: *-сы/-се*),
• категория притяжательности (PssAtr `attributive possesive`: *-дыкы/-деке*),
• категория уменьшительно-ласкательности (Dimin `diminutive`: *-кай/- кәй*),
• категория уподобления (Comp `comparison`, comp1: *-дай/ -дәй;* comp2: *-са/- сә*),
• категория атрибутивного локатива (LocAtr `attributive locative`: *-тағы/-тәге*),
• категория обладательности (CmtAtr `attributive comitate`: *-лы/-ле*),
• категория лишительности (Abs `abessive`: *-һыҙ/-һеҙ*),
• категория предельности (Term `terminative`: *-ғаса/-гәсә*),
• категория сравнительной степени (DgCom `degrees of comparison` аффиксом *-рақ/-рәк*).

Для категорий принадлежности и сказуемости выделяется подкатегория лица: p1, p2, p3.

Глагольные морфологические признаки включают показатели 11 категорий:
• Категория вопросительности (Q `question`: *-ме/-ме*).
• Категория неопределенности (Indf `indefinite`:*-дыр/-дер*).
• Категория усиления (Int `intensifying`: *-сы/-се*).
• Категория отрицания (Neg `negative`: *-ма/-мә*).
• Категория наклонения (Ind `indicative` изъявительное, Cond `conditional` условное, Opt `optative `желательное, Imp `imperative `повелительное, Intnl `intentional` намерения).
• Категория деепричастия (Ger `gerund`: Ger1: *-еп/-ып*; Ger2: *-ғас/-гәс*; Ger3: *-ғанса/-гәнсә*; Ger4: *-ғансы/-гәнсе*).
• Категория причастия (Ptcp `participle`: Prs: *-ыусы/-еүсе*; Pst: *-ған/-гән*, Fut1: *-асақ/-әсәк*; Fut2: *-ыр/-ер*).
• Категория имени действия (Act: *-ыу/-еү*).
• Категория инфинитива (Inf `infinitive`: *-рға/-ргә*).
• Категория хабитуалиса (Hab `habitualis`: *-сан/-сән*).
• Категория образования абстрактных субстантивов (Abst `abstractness`: - *лык/-лек*).

В корпусе размечаются следующие подкатегории глагольных форм:
• Времена (Prs `present` настоящее время, Fut `future` будущее время:
FutIndf `Future indefinite tense` будущее неопределенное время, FutDef `Future definite tense` будущее определенное время, Pst, прошедшее время, PstIndf `Past indefinite tense` прошедшее неопределенное время, PstDef `Past definite tense` прошедшее определенное время, PqpfDef `Plusquamperfect definite tense` предпрошедшее определенное время - ғайным/-гәйнем).
• Подкатегория лица (p: 1-3).
• Подкатегория числа (sg, pl).

Рис. 3. Морфологическая разметка текстов афористических жанров башкирского фольклора.



4. Заключение.

Таким образом, создание электронной базы текстов афористических жанров башкирского фольклора, интегрированная в корпус фольклора дает возможность пользователю исследовать язык текстов афористических жанров башкирского фольклора, получить лингвистические данные относительно морфологической информации представленных текстов. Преимуществом корпуса является и доступ к полному тексту произведений, т.к. именно текст в большинстве случаев является главной единицей анализа.

Данный корпус ориентирован для исследователей с разными интересами и задачами – лингвистам, фольклористам, журналистам, преподавателям, учащимся.

Литература:

1. Башҡортса-русса фразеологик һүҙлек. – Өфө: Башҡ. Китап нәшр., 1973. 168 б.
2. Башҡорт халыҡ ижады. I т. Йола фольклоры. Өфө, 1995.
3. Башҡорт халыҡ ижады. IX т. Йомаҡтар. Өфө, 2007.
4. Башҡорт халыҡ ижады. X т. Мәҡәлдәр һәм әйтемдәр. Өфө, 2006.
5. Башкирско-англо-русский словарь адекватных пословиц и поговорок. Авт.- сост. – Ф.А. Надршина, Э.М. Зубаирова [Созинова]. – Уфа: Китап, 2002. – 160 с.
6. Бускунбаева Л.А., Сиразитдинов З.А., Ишмухаметова А.Ш., Ибрагимова А.Д., Мигранова Л.Г. Корпус текстов периодической печати на башкирском языке // Актуальные проблемы языков народов России: Материалы XII региональной конференции. Уфа, 2012. С. 139–141.
7. Бускунбаева Л.А., Сиразитдинов З.А., Ибрагимова А.Д., Мигранова Л.Г., Полянин А.И. Башкирские языковые корпусы в Лаборатории лингвистики и информационных технологий // Урал и просторы Евразии сквозь века и тысячелетия. Уфа, 2013. С. 135-140.

8. Бускунбаева Л.А., Сиразитдинов З. А. О проблемах создания национального корпуса башкирского языка // Материалы Международной научно-теоретической конференции «Современное казахское языкознание: актуальные вопросы прикладной лингвистики», посвященная 75-летнему юбилею известного ученого профессора Жубанова Аскара Кудайбергенулы. Алматы, 2012. С. 54–58.
9. Бускунбаева Л.А., Сиразитдинов З.А. Система разметок в национальном корпусе башкирского языка // Языки меньшинств в компьютерных технологиях: опыт, задачи и перспективы. Йошкар-Ола, 2011, С. 45-51
10. Духовное наследие: фольклор свердловских башкир. – Уфа: ООО «Деловая династия», 2008. – С. 207-247. (на башк. яз.).
11. Корпусы Института этнологии и антропологии РАН: http://corpora.iea.ras.ru/corpora/ (дата обращения: 07.06.2017).
12. Корпус нганасанских фольклорных текстов: http://www.ilingran. ru/gusev/Nganasan/texts/index.php (дата обращения: 07.06.2017).
13. Куканова В.В. Фольклорный подкорпус: проблемы, структура и перспективы использования//Участие калмыков в укреплении российской государственности: материалы региональной научно-практической конференции Элиста: КИГИ РАН, 2012. С.193-198
14. Надршина Ф.А. Русско-башкирский словарь пословиц эквивалентов. – Уфа: Китап, 2008. – 196 с.
15. Николаев Д.С. Создание электронного корпуса фольклорных текстов на русском языке// V социологическая Грушинская конференция, 13 марта 2015 https://wciom.ru/fileadmin/file/nauka/grusha2015/s2_6/Nikolaev.pdf. (дата обращения: 07.06.2017).
16. Нәҙершина Ф.А. Халыҡ һүҙе. Өфө, 1983.– 160 б.
17. Салчак А.Я. Электронный корпус текстов тувинского языка // Новые исследования Тувы (Электронный журнал). №3, 2012 URL: https://www.tuva.asia/journal/issue_15/ (дата обращения: 07.06.2017).
18. Сиразитдинов З.А., Бускунбаева Л.А., Барлыбаева А.Д., Ишмухаметова А.Ш. К разработке корпуса прозаических текстов башкирского языка с 1917 по 1940-е годы // Этногенез. История. Культура. I Юсуповские чтения: Материалы Международной научной конференции, посвященной памяти Рината Мухаметовича Юсупова. Уфа, 2011. С. 269–274.
19. Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Ибрагимова А.Д. Информационные системы и базы данных башкирского языка. Уфа: Книжная палата РБ, 2013. – 116 с.
20. Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Ибрагимова А.Д. О создании корпуса башкирского фольклора // Урал-Алтай: через века в будущее: Материалы VI Всероссийской тюркологической конференции (с международным участием). Уфа, 2014. С. 86–89.
21. Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш. Лингвистические корпусы Лаборатории лингвистики и информационных технологий ИИЯЛ УНЦ РАН // Проблемы изучения национальных литератур: Материалы международной научной конференции. Махачкала, 2015. С. 658-664.
22. Сиразитдинов З.А., Полянин А.И. Об опыте разработки интегрированной корпусной системы на базе СУДБ Оракл // Труды казанской школы по компьютерной и когнитивной лингвистике TEL–2014. Казань, 2014. С.85–88.
23. Sirazitdinov Z.A. The corpora of the bashkir language // Turklang 14 Proceedings of the International Scientific Conference. Istanbul, 2014. P. 125–129.