ЯЗЫКОЗНАНИЕ И ЛИТЕРАТУРОВЕДЕНИЕ

УДК 004.658:801.3=512.141

КОРПУСНЫЕ ПРОЕКТЫ ЛАБОРАТОРИИ ЛИНГВИСТИКИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ ИИЯЛ УНЦ РАН

© З.А. Сиразитдинов

Рассматривается общее состояние корпусной лингвистики в зарубежной и отечественной лингвистике и вопросы разработки корпусов в Институте истории, языка и литературы УНЦ РАН. Анализируется деятельность лаборатории лингвистики и информационных технологий в рассматриваемой области. Описываются предлагаемые методы создания корпусов прозаических и публицистических текстов башкирского языка, ставится задача на перспективу.

Ключевые слова: корпусная лингвистика, башкирский язык, информационные системы, прикладная лингстистика.

Зародившееся в 60-х гг. прошлого века направление в зарубежном языкознании, связанное с компьютерной обработкой больших объемов текстов, сформировалось в новое быстро растущее направление филологии – корпусная лингвистика – «со своими традициями, признанными авторитетами, научными центрами, методами и проблематикой» [1]. Данному направлению сегодня во всем мире уделяется значительное внимание. Объектом нового филологического направления являются речевые материалы, реализованные в виде как письменных текстов, так и устных (фонетических) массивов данных. Корпусная лингвистика занимается созданием общих унифицированных принципов представления таких сверх-больших массивов языковых данных (корпусов), непосредственным созданием самих корпусов и выполнением конкретных экспериментальных лингвистических исследований на базе этих данных [2–3]. Данное направление лингвистики является приоритетным и в отечественной филологии. Так, если в «Плане фундаментальных исследований Российской академии наук на период 2006–2010 гг.» был раздел 9.2.3., касающийся создания электронного корпуса текстов русского языка, то в «Плане фундаментальных исследований Российской академии наук на период 2011–2025 гг.» в разделе 9.(б) ставится научная задача создания электронных корпусов текстов языков народов Российской Федерации [3]. Научный фонд РФФИ отдельно выделил корпусные исследования в своем классификаторе (06.4.20, Корпусно-ориентированные исследования) [4].

На сегодня в мире насчитываются более тысячи корпусов, количество их растет экспоненциально. Первый корпус был разработан в 60-х гг.. Это Брауновский корпус американского варианта современного английского языка, создававшийся в Брауновском университете в 1962–1963 гг. Объем корпуса около 1 млн словоупотреблений. В начале 2000-х гг. был создан корпус русского языка, на сегодня его объем составляет более 500 млн словоупотреблений.

Вся совокупность имеющихся корпусов весьма различна, поскольку, как было отмечено выше, объектом самой корпусной лингвистики являются многообразие речевых и письменных материалов языка. Так, по английскому, немецкому, китайскому, японскому, турецкому, эстонскому, русскому, польскому языкам реализованы речевые корпусы, содержащие как мультимедийные данные, так и транскрипции речи [5–11]. На стадии создания находятся корпусы и по другим языкам [12–13].

СИРАЗИТДИНОВ Зиннур Амирович – к.филол.н., Институт истории, языка и литературы УНЦ РАН, e-mail: sazin11@mail.ru

Но наибольшее количество корпусов составлено по письменным текстам. От поставленных целей и задач создания эти корпусы можно по разному классифицировать. Если корпус создается по текстам одного языка, то такой корпус является одноязычным. По объему привлеченных текстовых материалов среди них выделяются корпусы немецкого (DeReKo, 5,4 млрд слов) [14], английского (BNC, 100 млн слов) [15], американского варианта английского (450 млн слов) [16], китайского (LIVAC Synchronous Corpus, 1 млрд слов) [17], венгерского (100 млн слов) [18], испанского (100 млн слов) [19], итальянского (100 млн слов) [20], чешского (200 млн слов) [21], русского (НКРЯ, 500 тыс. слов) [22] языков. Если же создаются корпусы текстов переведенных на разные языки, то возникают многоязычные или, по другому, параллельные корпусы. Примерами таких корпусов являются польско-украинский, польско-русский, черногорско-английский, нидерландско-французский, японско-английский и другие параллельные корпусы [23–27]. Такие корпусы используются для сравнительно-соспоставительных исследований. Но в последнее время параллельные корпусы нашли практическое применение в разработках систем статистического перевода, зачинателем которого является компания Google. Одним из ярких примеров такого использования является параллельный корпус слушаний Европарламента, включающий тексты на 21 европейском языке [28].

В зависимости от стилистической принадлежности текстов выделяются художественные, научные [29–30], публицистические [31–33], драматургические, поэтические корпусы [34].

Текстовые корпусы также различаются по принципу отбора материала: выделяются полнотекстовые, когда в корпус попадают полные варианты печатных текстов, и фрагментно-текстовые. В последнем случае в корпус отбираются выборки из текстов. Объемы выборок и место расположения их в текстах каждый составитель определяет произвольно. Так, Брауновский корпус построен на базе выборок из 500 текстов, каждый из которых включает 2 000 словоупотреблений. Бирмингемский корпус английского языка и Основной корпус Национального корпуса русского языка являются представителями полнотекстового корпуса [35, с. 66; 22].

Для решения различных лингвистических задач мало лишь наличия массива текстов. Требуется также, чтобы сами тексты содержали в себе дополнительную лингвистическую информацию в виде специальных разметок, позволяющую использовать их для разных исследовательских и иных целей. В этой связи известный отечественный специалист в области составления корпусов, руководитель проекта Национального корпуса русского языка член-корр. РАН В.А. Плунгян даже подчеркивает, что «собственно, наука о корпусах … это прежде всего наука о том, как сделать хорошую разметку корпуса» [36, с. 6].

Составители корпусов по разному подходят к определению состава разметок, но большинство сходится в том, что разметки должны быть двух типов: экстралингвистические (метатекстовые) и лингвистические [37, с. 175–176]. К экстралингвистическим относится информация, которая паспортизирует сами тексты в целом и дает сведения об авторе (ФИО, год рождения автора, пол, образование и т.д.), информацию о тексте (название, год создания, год издания, жанр, тип текста, носитель текста: книга, журнал, электронное издание) и другие. Лингвистические разметки включают морфологические, синтаксические и семантические характеристики, относятся ко всем словоупотреблениям текста, поэтому некоторые авторы называют их лексическими разметками.

Для работы с размеченными текстами необходимо соответствующее программное сопровождение, позволяющее производить разнообразный поиск по корпусу, получать статистические данные. Размеченные тексты вместе с программным сопровождением образуют корпус в его полном понимании.

В создании корпуса трудоемким и сложным являются следующие этапы:

1) Подготовка электронных текстов. На данном этапе существующие печатные вари-

анты книг сканируются, редактируются и вводятся на электронные носители. Современные зарубежные корпусы создаются при поддержке крупных издательств, которые на безвозмездной основе передают предпечатные варианты текстов разработчикам корпусов.

2) Проведение разметки текстов. Степень трудоемкости данного этапа определяется уровнем развития таких разделов конкретного языка как компьютерная и математическая лингвистика. Если в языке проведены соответствующие исследования и составлена компьютерная модель, то возможны разработки средств автоматизации процесса. Первостепенной задачей в этом процессе является разработка автоматического морфологического анализатора языка. Далее следуют программы автоматического снятия омонимии, синтаксического и семантического анализа. Но даже в этом случае остается значительная доля ручной работы, поскольку не все языковые явления однозначно могут быть идентифицированы программными средствами.

Сейчас все крупные языки обзавелись своими национальными корпусами. К созданию корпусов приступили все остальные языки мира. Ведутся корпусные разработки и по языкам народов России: бурятского [38–39], калмыцкого [40–41], лезгинского [42] осетинского [43] и др. Отдельно отметим научные разработки и корпусные проекты по языкам тюркской группы, родственным башкирскому языку: казахский [44], татарский [45–46], тувинский [47–48], турецкий[491], шорский [50], хакасский [51].

Лингвистику XXI в. называют корпусной лингвистикой. При этом данное направление лингвистики активно влияет на все остальные направления языкознания, изменяет теоретические приоритеты и создает новые идеологии в понимании того, что же представляет собой язык [52; 7–8].

Исследователями также отмечается, что корпусы открывают перспективу для новых исследований не только в области лингвистики, но и в смежных областях: в литературоведении (для стилеметрических исследований, определения нормативности употребления языковых реалий), в общественных науках (изучение социальных объектов через язык, используя такие параметры текстов, как период, автор или жанр, семантический контент текстов), в информационно-технических разработках (создание автоматизированных систем машинного перевода, распознавание речи, информационный поиск).

Сегодня в Институте истории, языка и литературы УНЦ РАН активно осваиваются новые направления лингвистики прикладного характера, основывающиеся на накоплении лингвистических баз данных и компьютерной обработке. Есть первые результаты по экспериментальной фонетике, выполненные Л.К. Ишкильдиной [53]. Р.Н. Каримовой накапливается диалектная текстологическая и речевая база [54–55], разработан машинный фонд башкирского языка [56]. З.А. Сиразитдиновым и Л.Г. Мигрaновой составляется база терминологических данных [57], полным ходом идет работа и по корпусной лингвистике.

Работа по корпусу башкирского языка осуществляется сотрудниками лаборатории лингвистики и информационных технологий ИИЯЛ УНЦ РАН (З.А. Сиразитдинов, Л.А. Бускунбаева, А.Ш. Ишмухаметова, А.Д. Ибрагимова, Л.Г. Мигранова, А.И. Полянин) в двух направлениях: а) корпус прозаических текстов; б) корпус публицистических текстов.

Первое направление разрабатывается по гранту РФФИ «Разработка корпуса прозаических текстов башкирского языка», № 11-06-97001-р_поволжье_а. Начало работы 2011 г., окончание – 2013 г.

Второе направление осуществляется в рамках программы Президиума РАН «Корпусная лингвистика. Создание и развитие корпусных ресурсов по языкам народов России». Сроки реализации 2012–2014 гг. [58].

На сегодня по корпусу прозаических текстов разработаны системы экстралингвистических и лингвистических помет для разметки, создана программа автоматического морфологического анализа, подготовлены и автоматически размечены тексты 773 произведений более 70 авторов общим объемом

порядка 10 миллионов (10829086) словоформ, запущен проект поисковой системы в сети [59]. Сейчас идет отладка и оптимизация работы корпуса, ведется работа по оцифровке новых текстов. К концу года намечается доведение объема корпуса до 20 млн словоформ и запуск самого корпуса в сети Интернет на сервере Института со своим доменным именем. Проект корпуса прозаических текстов полностью разработан на базе СУБД Оракл на платформе Unicode [http://mfbl.ru/bashkorp/korpusp]. Для работы с корпусом пользователь может установить башкирскую раскладку клавиатуры средствами системы (ОС Vista, Seven), установить программу Хамелеон 8.0 (для ОС 98, ME, 2000, XP) или воспользоваться виртуальной клавиатурой самого корпуса.

По второму направлению подготовлены тексты республиканских газет и журналов общим объемом в 5 млн словоформ. Идет работа по автоматической морфологической разметке. Корпус будет выставлен к концу текущего года.

Система экстралингвистических разметок публицистического корпуса включает название прессы, год, месяц и день выхода, название статьи, автора. Все тексты размечены по тематике и жанру. Для рассматриваемого корпуса выделены следующие тематики и жанры.

Тематика: политическая и социальная жизнь (политика, право, философия); экономика (производство, строительство, бизнес, финансы, коммерция); сельское хозяйство; искусство, культура и литература; наука и техника; образование; природа, путешествие; частная жизнь; спорт; религия; психология; медицина; красота и здоровье.

Жанры текстов: интервью, беседа, статья, очерк, репортаж, обозрение, советы, письма, обзор печати (новости из других источников), поздравления, художественно-публицистические жанры (эссе, фельетон, рассказ, стихи, эпиграммы), рецензия.

По корпусу же прозаических текстов нами выделяются только авторы, названия произведений, год издания/завершения работы над произведением. Так, например, разрабатывемые корпусы текстов башкирского языка по классификации В.П. Захарова [2; 12–13] относятся к следующим типам (см. табл.):

Таблица

по типу языковых данных	письменный
по параллельности	одноязычный
по критерию литературности	литературный
по жанру	литературный, публицистический
по доступности	свободный доступ
по разметке	размеченный
по характеру разметки	морфологический, семантический
объем текстов	полнотекстовый

Система морфологической разметки обоих корпусов ориентирована на представление всех регулярных словоизменительных грамматических форм, не всегда отражаемых и совпадающих с формами, принятыми в академической грамматике. Морфологическая информация башкирской словоформы в корпусе включает: а) частеречную характеристику; б) 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. Эти характеристики указываются в словаре основ.

Именные морфологические признаки включают показатели следующих 15 категорий: числа, падежа, принадлежности, сказуемости, вопросительности, неопределенности, усиления, притяжательности, уменьшительно-ласкательности, уподобления, атрибутивный локатив (дағы/тағы), обладательности, лишительности, предельности, сравнительной степени.

* Авторы выражают благодарность член-корреспонденту РАН А.В. Дыбо за ценные советы в разработке системы морфологических разметок башкирского языка.

Глагольные морфологические признаки включают показатели следующих 11 категорий: вопросительности, неопределенности, усиления, отрицания, наклонения, деепричастия, причастия, имени действия, инфинитива, хабитуалиса (**сан/-сән**: барыусан, үсегеүсән), образования абстрактных субстантивов (**-лыk/-лек**: етерлек, алырлыk).

В корпусе размечаются следующие подкатегории для глагольных форм: 1) времена (настоящее время, будущее время: будущее неопределенное время, будущее определенное время, прошедшее время: прошедшее неопределенное время, прошедшее определенное время, предпрошедшее определенное время **-ғайным/-гәйнем**); 2) подкатегория лица (1–3); 3) подкатегория числа (ед., мн.).

Для именных форм выделяются следующие подкатегории: 1) подкатегория лица (1–3); 2) подкатегория числа (ед., мн.).

Морфологический анализатор корпуса реализован на основе алгоритма последовательного вычленения из словоформы букв и сравнения остатка словоформы и вычлененного фрагмента со словарями основ и аффиксов башкирского языка.

Для правильной идентификации основы и аффиксов используются грамматические фильтры: 1. Фильтр соответствия фонетической структуры аффикса фонетической структуре основы 2. Фильтр соответствия сочетаний аффиксов нормативным правилам. Данный фильтр основывается на списках возможных моделей сочетания словоизменительных аффиксов башкирского языка, которые были нами ранее рассмотрены в одной из наших работ [60]. 3. Фильтр графической передачи на стыках фонем.

Словарь основ включает нарицательные и собственные слова. Нарицательная часть словаря основ состоит из 60 тыс. единиц, включает лексику литературного башкирского языка. Часть имен собственных словаря включает имена, фамилии, отчества, клички животных и людей, географические названия башкирского и русского языков, имеет объем порядка 20 тыс. единиц.

В словарях основ указаны части речи, типы нарушений сингармонизма и возможные остатки основ при словоизменительных процессах и прочие варианты.

Проект национального корпуса башкирского языка художественной прозы позволяет производить следующие операции:

- поиск словоформы,
- поиск леммы,
- поиск грамматических категорий словоизменений,
- поиск грамматических подкатегорий,
- поиск сочетаний грамматических категорий,
- поиск сочетаний грамматических подкатегорий,
- поиск сочетаний словоформ,
- поиск сочетаний лемм,
- выдача списка небашкирской лексики (вкраплений по языкам-источникам),
- построение частотного словаря словоформ,
- построение частотного словаря лемм.

Сегодня проект корпуса прозаических текстов активно используется сотрудниками отдела языкознания при составлении многотомного академического толкового словаря башкирского языка.

Перед коллективом лаборатории линвгистики и информационных технологий ИИЯЛ УНЦ РАН в 2013 г. стоят следующие задачи:

1) доведение объема корпуса до 20 млн словоупотреблений;

2) разработка системы выдачи статистических распределений по любому заданному пользователем подкорпусу;

3) разработка системы выдачи графических представлений статистических распределений.

Работа подготовлена при поддержке гранта РФФИ 11-06-97001-р_Поволжье_а «Разработка корпуса прозаических текстов башкирского языка».

ЛИТЕРАТУРА

1. Рыков В.В. Прагматически ориентированный корпус текстов // Тверской лингвистический мери-

диан. Тверь, 1999 (http://rykov-cl.narod.ru/t.html, дата обращения: 17.06.2013).

2. Захаров В.П. Корпусная лингвистика: учебно-методическое пособие. СПб., 2005. 48 с.

3. План фундаментальных исследований Российской академии наук на период 2011–2025 гг. URL: http://www.ras.ru/scientificactivity/plan2025.aspx (дата обращения: 17.06.2013).

4. Классификатор РФФИ. URL:http://scs.viniti.ru/rubtree/main.aspx?tree=RFFI&cod=06 (дата обращения: 17.06.2013).

5. LDC Top Ten Corpora (мультимедийные корпусы английского языка). URL: http://www.ldc.upenn.edu/Catalog/topten.jsp (дата обращения: 17.06.2013).

6. Chinese Broadcast Conversation Speech (мультимедийный корпус китайского языка). URL: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2013S04 (дата обращения: 17.06.2013).

7. The Corpus of Spontaneous Japanese (мультимедийный корпус японского языка). URL: http://www.ninjal.ac. jp/products-k/katsudo/seika/corpus/public/index.html (дата обращения 17.06.2013).

8. The Spoken Turkish Corpus (мультемидийный корпус разговорного турецкого языка). URL: http://stc.org.tr (дата обращения: 17.06.2013).

9. Фонетический корпус спонтанной эстонской речи. URL: http://www.murre.ut.ee/phonetic-corpus (дата обращения: 17.06.2013).

10. Фонетический немецкого разговорного языка. URL: http://dsav-wiss.ids-mannheim.de/korpora/pf/pf _doku.htm (дата обращения: 17.06.2013).

11. Фонетические корпусы русского и польского языков URL: http://www.voicemethods.com/new/databases/corpuses.php3 (дата обращения: 17.06.2013).

12. Людовик Т.В., Робейко В.В., Пилипенко В.В. Автоматическое распознавание спонтанной украинской речи (на материале акустического корпуса украинской эфирной речи) // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). Вып. 10 (17). М.: Изд-во РГГУ, 2011. С. 478–489.

13. Крючкова О. Ю., Гольдин В. Е. Корпус русской диалектной речи: концепция и параметры оценки / Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). М.: Изд-во РГГУ, 2010. С. 359–368.

14. Das Deutsche Referenzkorpus (DeReKo) URL: http://www.ids-mannheim.de/kl/projekte/korpora (дата обращения: 17.06.2013).

15. British National Corpus (BNC). URL:http://www.natcorp.ox.ac.uk (дата обращения: 17.06.2013).

16. The corpus of contemporary american english (COCA). URL:http://corpus.byu.edu/coca (дата обращения: 17.06.2013).

17. Корпус китайского языка. (LIVAC Synchronous Corpus). URL: http://www.rcl.cityu.edu.hk/livac (дата обращения: 17.06.2013).

18. Magyar Nemzeti Szövegtár (корпус венгерского языка). URL: http://corpus.nytud.hu/mnsz (дата обращения: 17.06.2013).

19. Corpus del español (корпус испанского языка). URL:http://www.corpusdelespanol.org (дата обращения: 17.06.2013).

20. Corpus di riferimento della lingua italiana scritta contemporanea («CoLFIS») (корпус итальянского языка) URL: http://www.ge.ilc.cnr.it/dizionari.php (дата обращения: 17.06.2013).

21. Český národni korpus (ČNK) (чешский национальный корпус). URL: http://ucnk.ff.cuni.cz (дата обращения: 17.06.2013).

22. Национальный корпус русского языка. URL: http://www.ruscorpora.ru (дата обращения: 17.06.2013).

23. Польско-украинский параллельный корпус. URL:http: //www.domeczek.pl/~polukr/ index.php?option=welcome (дата обращения: 17.06.2013).

24. Польско-русский параллельный корпус. URL: http://pol-ros.polon.uw.edu.pl (дата обращения: 17.06.2013).

25. Englesko-crnogorski paralelni korpus (черногорско-английский параллельный корпус). URL: http://www.eiprevod.gov.me/korpus (дата обращения: 17.06.2013).

26. Dutch-French Parallel Corpus (DPC) (нидерландско-французский параллельный корпус). URL: http://dpc.inl.nl/indexd.php (дата обращения: 17.06.2013).

27. Japanese-English Parallel Corpus (японско-английский параллельный корпус). URL: http://www.manythings.org/corpus (дата обращения: 17.06.2013).

28. European Parliament Proceedings Parallel Corpus 1996–2011 (параллельный корпус слушаний

Европарламента). URL: http://www.statmt.org/europarl (дата обращения: 17.06.2013).

29. Corpus Albaruthenicum (корпус научных белорусских текстов). URL: http://grid.bntu.by/corpus/ (дата обращения: 17.06.2013).

30. Zientzia eta Teknologiaren Corpusa *(н*аучно-технический баскский корпус). URL: http://www.ztcorpusa.net/cgi-bin/kontsulta.py (дата обращения: 17.06.2013).

31. Корпус русских публицистических текстов второй половины XIX века. URL: http://smalt.karelia.ru/corpus/index.phtml (дата обращения: 17.06.2013).

32. Компьютерный корпус текстов русских газет конца XX века. URL: http://www.philol.msu.ru/~lex/corpus/ (дата обращения: 17.06.2013).

33. Romanian corpus (корпус румынской прессы). URL: http://corp.hum.sdu.dk/cqp.ro.html (дата обращения: 17.06.2013).

34. Поэтический подкорпус НКРЯ. URL: http://www.ruscorpora.ru/search-poetic.html (дата обращения: 17.06.2013).

35. Баранов А.Н. Введение в прикладную лингвистику: учебное пособие. М.: Эдиториал УРСС, 2001. 360 с.

36. Плунгян В.А. Зачем нужен Национальный корпус русского языка? Неформальное введение // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 6–20.

37. Поляков А.Е. Технология подготовки информации в национальном корпусе русского языка // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005. С. 175–192.

38. 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. СПб., 2008. С. 24–30.

39. Корпус бурятского языка. URL: http://web-corpora.net/BuryatCorpus/search/?interface_language=ru (дата обращения: 17.06.2013).

40. Куканова В. В. Архитектура метаописания в национальном корпусе калмыцкого языка // Вестник Калмыцкого института гуманитарных исследований РАН. 2011. № 1. С. 139–145.

41. Корпус калмыцкого языка. URL:http://web-corpora.net/KalmykCorpus/search/?interface_language=ru (дата обращения: 17.06.2013).

42. Корпус лезгинского языка. URL:http://www.dag-languages.org/LezgianCorpus/search/ (дата обращения: 17.06.2013).

43. Корпус осетинского языка. URL:http://www.ossetic-studies.org/iron-corpus/search/index.php?interface_language=ru. (дата обращения: 17.06.2013).

44. Жұбанов А.Қ. Қазақ тілінің аннотацияланған мәтіндер корпусындағы етесті сөздерге лексик-морфологиялық белгі-код (белгіленім) қоюдың алғышарттары. «Тілтаным», 2012. № 1, 18–25 б. (Журнал Института языкознания им. А. Байтурсынова, Казахстан, Алматы).

45. Сулейманов Д.Ш., Хакимов Б.Э., Гильмуллин Р.А. Корпус татарского языка: концептуальные и лингвистические аспекты // Вестник Татарского государственного гуманитарно-педагогического университета. № 4(26), 2011. С. 211–216.

46. Письменный корпус татарского языка. URL:http://corpus.tatfolk.ru (дата обращения: 17.06.2013).

47. Салчак А. Я. 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: 17.06.2013).

48. Проект тувинского корпуса. URL: http://www.tuvancorpus.ru (дата обращения: 17.06.2013).

49. Sözlü Türkçe Derlemi *(*корпус разговорного турецкого языка). URL: http://std.metu.edu.tr (дата обращения: 17.06.2013).

50. Электронный корпус шорских текстов. URL: http://shoriya.ngpi.rdtc.ru (дата обращения: 17.06.2013).

51. Шеймович А.В. Морфологическая разметка корпуса хакасского языка // Российская тюркология. 2011. № 2 (5). С. 48–61.

52. Плунгян В.А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. 2008. № 16 (2). С. 7–20.

53. Ишкильдина Л.К., Уртегешев Н.С. Фонема [w] башкирского языка: функционирование, история развития, артикуляторные характеристики (по данным томографирования) // Тумашевские чтения: актуальные проблемы тюркологии: мат-лы IV Всероссийской научно-практической конференции. Тюмень: Печатник, 2010. С. 442–446.

54. Каримова Р.Н. Текстологический электронный корпус башкирских говоров // Урал-Алтай: через века в будущее: мат-лы IV Всероссийской научной конференции, посвященной III Всемирному курултаю башкир. Уфа, 2010. С. 189–191 (на башк. яз.).

55. Каримова Р.Н. Электронный фонд экспедиционных аудиозаписей // Урал-Алтай: через века в будущее: мат-лы IV Всероссийской научной конференции, посвященной III Всемирному курултаю башкир. Уфа, 2010. С. 162–163.

56. Сиразитдинов З.А., Максутов А.Д., Полянин А.И., Бускунбаева Л.А. Информационная лингвистическая система «Машинный фонд башкирского языка» // Урал-Алтай: через века в будущее: мат-лы IV Всероссийской научной конференции, посвященной III Всемирному курултаю башкир (25–27 марта 2010 г.). Уфа, 2010. Т. 1. С. 286–290.

57. Сиразитдинов З.А., Мигранова Л.Г., Ишмухаметова А.Ш., Ибрагимова А.Д., Бускунбаева Л.А. К созданию терминологического банка данных башкирского языка // Урал-Алтай: через века в будущее: мат-лы V Всероссийской конференции, посвященной 80-летию учреждения РАН ИИЯЛ УНЦ РАН (21–22 июня, 2012 г.). ИИЯЛ УНЦ РАН. 2012. Уфа, 2012. С. 111–114.

58. Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Ибрагимова А.Д., Мигранова Л.Г. Корпус текстов периодической печати на башкирском языке // Актуальные проблемы диалектологии языков народов России: мат-лы XII региональной конференции. Уфа, 2012. С. 139–141.

59. Сиразитдинов З.А., Ибрагимова А.Д., Ишмухаметова А.Ш., Полянин А.И. О пилотном проекте национального корпуса прозаических текстов башкирского языка // Урал-Алтай: через века в будущее: мат-лы V Всероссийской конференции, посвященной 80-летию учреждения РАН ИИЯЛ УНЦ РАН (21–22 июня, 2012 г.), ИИЯЛ УНЦ РАН. 2012. Уфа, 2012. С. 108–111.

60. Сиразитдинов З.А. Моделирование грамматики башкирского языка. Словоизменительная система. Уфа: Гилем, 2006. 160 с.

CORPUS-BASED PROJECTS OF THE LABORATORY OF LINGUISTICS AND INFORMATION TECHNOLOGY (IHLL USC RAS)

© Z.A. Sirazitdinov

Institute of History, Language and Literature, USC RAS, Ufa, Russin Federation

The article discusses the general status of corpus linguistics in Russia and abroad and the issues of corpus development at the Institute of History, Language and Literature, Ufa Scientific Centre, RAS. It analyzes the work of the Laboratory of Linguistics and Information Technology in the area in question, describes the proposed methods for creating corpora of prosaic and publicistic texts in the Bashkir language and sets the task for the future.

Key words: corpus linguistics, the Bashkir language, information systems, applied linguistics.