Блинова О. В. Будешь задумываться / что говорить / а что не говорить: парадокс наблюдателя и речевой контроль в ОРД. Режим доступа: https://www.slm.uni-hamburg.de/slavistik/forschung/veranstaltungen/symposium-sprachvariation/downloads-symposium-april-2016/blinova.pdf

Бускунбаева Л. А. Закономерности речевой экономии и их отражение в башкирском языке. Уфа, 2008.

Звуковой корпус как материал для анализа русской речи. Ч. 1: Чтение. Пересказ. Описание / Под ред. Н. В. Богдановой-Бегларян. СПб., 2013.

Кибрик А. А., Подлесская В. И. К созданию корпусов устной русской речи: принципы транскрибирования // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2003. № 10. С. 5–13.

Куканова В. В., Бембеев Е. В., Убушаев Н. Н., Манджиева Б. Б. Устные тексты на калмыцком языке: запись и расшифровка // Вестник калмыцкого университета. 2013. № 3 (19). С. 56–64.

Сиразитдинов З. А., Бускунбаева Л. А., Ишмухаметова А. Ш., Шамсутдинова Г. Г. Проблемы разработки диалектного корпуса башкирского языка // Вопросы диалектологии: Международный научный журнал. 2018. № 1–2. С. 97–107.

Степанова С. Б., Асиновский А. С., Богданова Н. В., Русакова М. В., Шерстинова Т. Ю. Звуковой корпус русского языка повседневного общения «один речевой день»: концепция и состояние формирования // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. «Диалог». Вып. 7 (14). М., 2008. С. 488–495.

Щерба Л. В. Фонетика французского языка. Очерк французского произношения в сравнении с русским: Пособ. для студент. факульт. иностр. языков. М., 1953.

УДК 811.512.145

Сайхунов Мансур Равхатович
Институт языка, литературы и искусства им. Г. Ибрагимова АН РТ
г. Казань, Республика Татарстан
Хусаинов Рустем Рафаэлевич
Компания «GDC»
г. Казань, Республика Татарстан
Ибрагимов Тавзих Ибрагимович
ФГАОУ ВО «Казанский (Приволжский) федеральный университет»
г. Казань, Республика Татарстан

СЛОЖНОСТИ ПРИ СОЗДАНИИ ТЕКСТОВОГО КОРПУСА ОБЪЕМОМ БОЛЕЕ 400 МЛН ТОКЕНОВ

Аннотация: В статье рассматривается ряд вопросов, связанных с различными этапами создания корпуса текстов: сбор и обработка материала, критерии группировки текстов, разметка и т.д. Авторы обращают внимание на такие технические нюансы, как выбор корпусного менеджера, подбор хостинга для проекта, создание на базе корпуса дополнительных сервисов и др. В ходе исследования были выработаны определенные рекомендации, которые могут быть полезны при создании других текстовых корпусов.

Ключевые слова: татарский язык, языковой корпус, морфологическая разметка, корпусный менеджер.

Создание корпусов национальных языков является актуальной задачей современной лингвистики. Отечественными исследователями активно разрабатываются корпусы русского, башкирского, калмыцкого, коми и многих других языков.

Письменный корпус татарского языка [ПКТЯ], создаваемый авторами статьи, не финансируется какими-либо научными фондами или организациями. Все работы над данным проектом ведутся исключительно в свободное время участников.

Начальные шаги по созданию корпуса, такие как обдумывание структуры, поисковой системы, списка включаемых материалов, относятся к 2010 г. Первая рабочая версия, содержащая базу из 45 млн словоупотреблений и около 60 источников, была опубликована в интернете в начале 2012 г.

Вторая версия корпуса, характеризующаяся увеличением объема корпуса до 116 млн словоупотреблений при 2750 источниках и введением морфологической разметки, вышла в 2014 г.

Последняя третья версия корпуса общим объемом 356 млн слов (430 млн токенов) из 16786 источников была выложена в сеть интернет в конце 2018 г.

При рассмотрении этих цифр, может создаться впечатление, что разница между различными версиями сводится лишь к добавлению определенной порции новых текстов. Оглядываясь назад, мы видим, что в технологическом плане понятие «языковой корпус» гораздо шире, чем простая коллекция текстов с интегрированной системой поиска. Процесс создания и развития корпуса того или иного языка включает множество этапов, на которых порой объединяются различные технологии и подходы, часто несовместимые и нестандартизированные в силу новизны. В процессе реализации данного проекта должно быть решено множество задач, зависящих от наличия финансовых, технических и человеческих ресурсов, доступа к специализированным программным средствам, обладания необходимой квалификацией и опытом, учета особенностей языка и имеющегося материала. В течение последних девяти лет в ходе работы над Письменным корпусом татарского языка с каждой новой версией авторы сталкивались с теми или иными проблемами.

1. Поиск новых источников и обновление имеющихся

Важнейшей частью корпуса является коллекция текстов, поэтому данный этап имеет наибольшее значение. С увеличением объема становится сложнее найти новые источники для включения в корпус, соблюдая баланс в репрезентативности. Особенно это касается малых языков, т.к. они представлены в сети интернет в ограниченном объеме. Определение новых ресурсов для Письменного корпуса татарского языка, в основном, совершается через поиск наиболее частотных татарских слов в популярных поисковых системах с ручным просмотром и отбором найденных результатов: новостные сайты, веб-порталы государственных ведомств и учреждений, онлайн коллекции различной тематики, персональные сайты школьных учителей и др. Списки новых книг для последующей покупки и сканирования составляются из представленных в книжных магазинах и на сайтах издательств.

2. Сбор текстов из отобранных источников

После подготовки списка включаемых в корпус источников, начинается этап сбора текстов. Данная работа ведется по нескольким направлениям.

а) Одно из них это так называемый «краулинг», т.е. автоматическое или полуавтоматическое скачивание определенных сайтов. Для этого могут быть использованы различные программные продукты, например, от самых простых, как Wget [Wget] или Curl [Curl], до специализированных сложных систем вроде Heritrix [Heritrix]. Обычно целиком скачанные веб-сайты состоят из сильно варьирующихся по качеству текстов, поэтому мало подходят в качестве авторитетных источников примеров употребления тех или иных языковых явлений. Однако подобный материал представляет большую ценность для получения общей статистической информации о языке, его богатстве и возможностях, тенденциях развития.

б) В другом случае, как, например, поиск конкретных примеров для тех или иных языковых явлений, требует наличия в корпусе большого числа изолированных текстов различного жанра: художественные произведения прозаического и поэтического характера, научная литература (в случае миноритарных языков, в основном, гуманитарного профиля), тексты религиозного содержания, научно-популярные, публицистические труды и др. Подобные тексты позволяют исследователям находить нужную для них информацию в проверенных источниках, на которые можно авторитетно ссылаться, например, в словарях, грамматиках и монографиях. Данные произведения в основном собираются в ручном или полуавтоматическом режиме: личное получение текстов из рук самих писателей, их родственников и других близких им людей, через сотрудничество с издательствами и библиотеками, поиск в интернете в открытых источниках.

3. Метаразметка и классификация собранных текстов

Метаразметка представляет собой присваивание текстам дополнительной описательной информации. На сегодняшний день имеются спецификации Text Encoding Initiative (TEI), а также рекомендации EAGLES (Expert Advisory Group on Language Engineering Standards) оформленные в качестве предлагаемого корпусного стандарта CES (Corpus Encoding Standard) для унификации представления метаданных в формате XML. Однако в Письменном корпусе татарского языка по историческим причинам применяется свой набор и именование описывающих документ данных, а вместо XML используется простой текстовый формат (plain text). В первую очередь это связано с тем, что тексты в таком формате гораздо удобнее автоматически обрабатывать различными UNIX утилитами и скриптами, как sort, uniq, wc, sed, awk и др. По мере совершенствования указанных стандартов и поддерживающих их программных продуктов считаем необходимым переход на их использование, чтобы корпусы могли быть определены как соответствующие стандарту (представление экстралингвистического и лексиче-

ского описания, общая архитектура) и могли быть прозрачными при обмене данными или объединении с другими проектами.

Тексты, собранные в Письменном корпусе татарского языка имеют экстралингвистическую разметку по следующим критериям:

а) Автор, т.е. сведения об авторе или группе авторов (имена и псевдонимы).

б) Название произведения, статьи, книги, журнала, газеты, сайта и т.д.

в) Время написания или издания.

г) Тип, т.е. стилевые особенности текста (поэзия, проза, публицистика, фольклор, официальный или научный текст).

д) Жанр (роман, повесть, рассказ, поэма, стихотворение, статья, поверье и др.).

е) Источник, т.е. библиографические данные книги, газеты или журнала, откуда данный текст был взят.

ж) URL адрес, а точнее доменное имя веб-сайта.

з) Дополнительные данные, т.е. какая-либо информация, не попавшая в другие категории (от кого получены данные, кто подготовил данный файл для включения в корпус, параметры кодирования, версия языка разметки и др.).

В связи с довольно быстрым ростом объема корпуса и нехваткой человеческих ресурсов не удается оперативно обрабатывать все новые материалы. Поэтому данная работа значительно отстает, и новые тексты изначально временно маркируются более общими тегами, например, вместо отдельных статей или произведений указывается целый сборник, вместо указания отдельных номеров журналов и газет, они объединяются под общим названием. В дальнейшем, по мере появления свободного времени, данные тексты постепенно получают полное метаописание.

4. Исключение повторяющихся текстов и предложений

Чем больше объем корпуса, тем сложнее избегать дублирования материалов. В Письменном корпусе для решения этой задачи авторы опираются на несколько принципов:

а) В предложениях, отличающихся лишь знаками препинания, трудно автоматически определить, которое из них орфографически корректно. Поэтому, следуя принципу «не навреди», сохраняются все варианты.

б) В рамках одного веб-сайта удаляются все дублирующиеся предложения. Это связано с тем, что веб-краулинг – довольно сложная процедура, в ходе которой неизбежно получение одних и тех же текстов или их частей повторно. К тому же веб-сайты, в основном, это новостные ресурсы, которые состоят, как правило, из множества однотипных новостей. Для них характерна слабая рецензия, что отражается на качестве и ценности материала.

в) Новостные статьи и предложения, которые повторяются на разных веб-сайтах, сохраняются в связи с тем, что трудно определить в автоматическом режиме, где первоисточник, а где перепубликация. Тем более новостные тексты очень малого объема могут дублироваться на разных ресурсах, но при этом быть оригинальными.

г) В художественных произведениях, научных трудах, законодательных актах и других уникальных текстах дублирующиеся предложения сохраняются, т.к. это отражает индивидуальный стиль автора или является характерной чертой определенной сферы применения. В дополнение к этому, подобные тексты, как правило, проходят строгую рецензию, что повышает их ценность.

д) Дублирование целых художественных, научных и других текстов или их больших частей должно быть минимизировано. Данная проблема лучше всего решается при максимальной категоризации текстов, где подобные случаи проявляются в процессе обработки.

5. Нормализация

Этап нормализации в Письменном корпусе, в первую очередь, связан с исправлением некорректно отображаемых татарских символов. Чаще всего это связано с использованием в старых текстах CP1251 кодировки, которая на сегодняшний день практически полностью вытеснена Юникодом (UTF-8).

Следующей проблемой, часто встречающейся в текстах, собранных с веб-сайтов, являются так называемые «сломанные» или «неправильные» Юникод (UTF-8) последовательности. На первых этапах они могут не представлять никаких проблем, но в дальнейшем приводят к сбоям при автоматической морфологической аннотации, импорте материала в систему управления базами данных (СУБД), работе поисковой системы и т.д. Данная проблема успешно решается стандартными перекодировщиками вроде iconv.

В нормализацию также входит удаление или преобразование табличных данных, удаление или конвертация таких непечатаемых символов, как переносы, неразрывные пробелы, приведение символов конца строк в единый формат. Сюда же можно отнести и приведение к некоему единообразному виду, например, следующих групп символов:

– В текстах Письменного корпуса татарского языка было выявлено употребление около десяти различных видов тире: ‐ ‑ ‒ – — ― ⸺ ⸻ - . Некоторые из этих символов визуально могут показаться идентичными, однако имеют различное кодовое представление и с точки зрения корпус-менеджера являются совершенно разными знаками.

– Использование в текстах кавычек также варьируется: « » » “ ” „ " ’ ‘ ′ ` ' .

6. Сегментация текста на предложения и абзацы

На этом этапе осуществляется сегментирование текста на его структурные составляющие: предложения, абзацы и другие единицы. В Письменном корпусе татарского языка не используется сегментация на абзацы. Выявление границ предложений зависит от конкретного языка в связи с тем, что должны учитываться:

– характерные для данного языка сокращения («һ.б.» ‘и другие’, «б.э.к.» ‘до нашей эры’) и аббревиатуры;

– инициалы («Г. Тукай, Ибраһимов Г. Г.») и другие нюансы.

7. Морфологическая аннотация

Для морфологической разметки Письменного корпуса используется система Apertium [Washington et al. 2014], которая является свободным программным обеспечением, имеющим полностью открытый исходный код. Apertium – это крупный международный проект по созданию открытой системы машинного перевода для большого числа языков. На сегодняшний день поддерживается более сотни языков, в том числе и языки малых народов России.

Автоматический анализ естественного языка, как правило, дает несколько вариантов аннотации для одной лексической единицы. Это называется грамматической омонимией. Снятие неоднозначности является важной задачей компьютерной лингвистики. На сегодняшний день практически невозможно снять омонимию ручным способом из-за огромного объема корпусов. Поэтому все больше исследуются способы автоматического решения данной проблемы, создаются специализированные программные комплексы, которые опираются на предопределенные правила, статистические данные или предобученные нейронные модели. Используемый в Письменном корпусе морфологический анализатор Apertium имеет основанную на наборе правил встроенную систему разрешения некоторых случаев грамматической неоднозначности. Еще одним преимуществом использования данного проекта является применение унифицированной для большого числа языков (особенно родственных) набора тегов, что позволяет создавать корпусы со структурно близкой морфологической разметкой для разных языков и облегчает проведение сопоставительных исследований.

Пример текста, аннотированного системой Apertium:

*^-/-<guio>$ ^**Әй**/Әй<ij>$^,/,<cm>$ ^**алар**/алар<n><sg><attr>$*
*^**авыллары**/авыл<n><pl><px3sp><nom>$ ^**белән**/белән<post>$*
*^**шулай**/шул<prn><dem><adv>$ ^**бит**/бит<mod_ass>$^./.<sent>$*

*^**Менә**/Менә<ij>$ ^**безнен**/*безнен$ ^**егетләр**/егет<n><pl><nom>$*
*^**бәйләнчек**/бәйләнчек<n><sg><nom>$ ^**түгел**/түгел<cnjcoo>$^,/,<cm>$*
*^**алар**/алар<prn><pers><p3><pl><nom>$ ^**белән**/белән<post>$*
*^**анлашып**/*анлашып$ ^**була**/бул<v><iv><pres><p3><sg>$^./.<sent>$*

*^**Якын җибәрмиләр**/Якын җибәр<v><tv><neg><pres><p3><pl>$*
*^**үзен**/үз<prn><ref><px3sp><acc>$^./.<sent>$*

*^-/-<guio>$ ^**Ә**/Ә<ij>$ ^**клубтан**/*клубтан$*
*^**кайткanымны**/кайт<v><tv><ger_past><px1sg><acc>$ ^**тагын**/тагын<adv>$*
*^**сагалап торса**/сагалап тор<v><tv><gna_cond><p3><sg>$^**?**/?<sent>$*

*^-/-<guio>$ ^**Борчылма**/Борчы<v><tv><pass><neg><imp><p2><sg>$^,/,<cm>$*
*^**нинди**/нинди<prn><itg><sim>$ ^**икәнеңне**/икән<n><sg><px2sg><acc>$*
*^**белгән**/бел<v><tv><gpr_past>$ ^**бит**/бит<mod_ass>$ ^**инде**/инде<adv>$^./.<sent>$*

*^**Чишмә**/Чишмә<np><top><attr>$ ^**суы**/су<n><sg><px3sp><nom>$*
*^**эчеп**/эч<v><tv><gna_perf>$^,/,<cm>$ ^**битләребезне**/бит<n><pl><px1pl><acc>$*

^**югач**/*ю<v><tv><gna_after>$^,/,<cm>$* ^**рәхәт**/*рәхәт<adj><advl>$*
^**булып**/*бул<v><iv><prc_perf>$* ^**китте**/*кит<vaux><ifi><p3><sg>$^./.<sent>$*

^**Җиләк**/*Җиләк<np><ant><f><nom>$* ^**җыюын**/*җыю<n><sg><px3sp><acc>$*
^**җыйдык та**/*җый<v><tv><ifi><p1><pl>+да<cnjcoo>$^,/,<cm>$*
^**өйгә**/*өй<n><sg><dat>$* ^**кайтач**/**кайтач$^,/,<cm>$*
^**чистартасы да**/*чистарт<v><tv><ger_fut3><px3sp><nom>+да<cnjcoo>$* ^**бар**/*бар<adj>$*
^**бит**/*бит<mod_ass>$* ^**әле**/*әле<adv>$* ^**аны**/*ул<prn><dem><acc>$^!/!<sent>$*

^**Әйдә**/*Әйдә<ij>$^,/,<cm>$* ^**яшь**/*яшь<adj>$*
^**бәрәнге**/*Бәрәнге<np><top><nom><err_orth>$*
^**пешердем**/*пешер<v><tv><ifi><p1><sg>$^,/,<cm>$*
^**катык та**/*катык<n><sg><sg><nom>+да<cnjcoo>$*
^**бар**/*бар<adj>+и<cop><aor><p3><sg>$^./.<sent>$*

^**Катык**/*Катык<n><sg><nom>$^**ашыйсым**/аша<v><tv><prc_vol><p1><sg>$*
^**килгәнне**/*кил<vaux><ger_past><acc>$^**каян**/каян<adv><itg>$*
^**белгәнсеңдер**/*бел<v><tv><past><p2><sg>+дыр<mod_ind>$^!/!<sent>$*

^**Үзе**/*Үз<prn><ref><px3sp><nom>$*
^**куе**/*куй<n><sg><px3sp><nom>+и<cop><aor><p3><sg>$^,/,<cm>$*
^**кызгылт-сары**/*кызгылт-сары<adj>$*
^**төстә/төс**<*n><sg><loc>+и<cop><aor><p3><sg>$^,/,<cm>$*
^**дерелдәп**/*дерелдә<v><tv><prc_perf>$*
^**тора**/*тор<vaux><pres><p3><sg>$^./.<sent>$*

^**Шундый**/*Шул<prn><dem><sim><nom>$* ^**шәп**/*шәп<adj>$*
^**егетне**/*егет<n><sg><sg><acc>$* ^**кулдан**/*кул<n><sg><sg><abl>$*
^**ычкындыралармыни**/**ычкындыралармыни$^?/?<sent>$*

8. Выбор корпусного менеджера и индексация текстовой базы

Корпус языка подразумевает наличие не только набора текстов, но и специализированной поисковой системы, включающей программные средства для поиска данных в корпусе, получения статистической информации и представления результатов пользователю в удобной форме [Захаров, Богданова 2013: 68]. Большинство корпусных менеджеров поддерживает обработку типовых запросов двух видов. Во-первых, это простой поиск по точному совпадению или по маске. Во-вторых, это так называемый язык корпусных запросов (Corpus Query Language), основанный на той или иной версии системы регулярных выражений: базовые или расширенные POSIX-совместимые, Perl-совместимые (PCRE).

Корпусные поисковые системы, будь то всемирно известные NoSketch Engine [NoSketch], CQP/CWB [The IMS] или разрабатываемый авторами статьи Fastmorph [Сайхунов и др. 2017: 382–385; 2018: 314–319], могут индексировать текстовые данные только определенного формата. Обычно это так называемый «вертикальный формат» с небольшими различиями в именовании и количестве используемых тегов. В данном формате каждое слово (и характеризующая его морфологическая, синтаксическая и любая другая аннотация) размещается в отдельной строке в виде колонок, а границы предложений, абзацев и текстов указываются специальными тегами. Метаданные размещаются внутри тега ***<doc>*** или ***<text>***, например:

<doc
id*="2263_Gayaz_Isxakyy_Kajul_chitek"*
author*="Гаяз Исхакый"*
title*="Кәҗүл читек"*
date*="1912"*
type*="PROSE"*
genre*="Хикәя"*
source*=""*
url*=""*
meta*=""*
lang*="Tatar"*
>

<s>

Минем	*мин*	*n*	*n:sg:sg:px1sg:nom*
әти	*әти*	*n*	*n:sg:sg:attr*
бүген	*бүген*	*adv*	*adv*

```
Казаннан    казан       np      np:top:abl
кайта       кайт        v       v:tv:pres:p3:sg
<g/>
.           .           sent    sent
</s>

 <s>
Улмы        ул          prn     prn:pers:p3:sg:nom:мы:qst
<g/>
?           ?           sent    sent
</s>

<s>
Ул          ул          prn     prn:pers:p3:sg:nom
миңа        мин         prn     prn:pers:p1:sg:dat
санлы       санлы       adj     adj
калач       калач       n       n:sg:nom
китерә      китер       v       v:tv:pres:p3:sg
<g/>
.           .           sent    sent
</s>

<s>
Аннары      аннары      adv     adv
миңа        мин         prn     prn:pers:p1:sg:dat
читек       читек       n       n:sg:nom
китерә      китер       v       v:tv:pres:p3:sg
<g/>
.           .           sent    sent
</s>

<s>
Өр          өр          v       v:tv:imp:p2:sg
<g/>
-           -           guio    guio
<g/>
яңа         яңа         adj     adj
читек       читек       n       n:sg:nom:u:cop:aor:p3:sg
<g/>
.           .           sent    sent
</s>

<s>
Үзе         үз          prn     prn:ref:px3sp:nom
кып-кызыл   кып-кызыл   adj     adj
төсле       төсле       adj     adj:u:cop:aor:p3:sg
<g/>
.           .           sent    sent
</s>
```

9. Поиск подходящего хостинга с учетом требований корпусной поисковой системы

С каждой новой версией Письменный корпус размещается на более ресурсоемком сервере. Сначала это был простой веб-хостинг с 7 Гб дискового пространства, где системные требования не превышали наличия интерпретатора PHP и СУБД MySQL. Следующая версия, где начала использоваться система поиска Fastmorph, а позже еще и корпус-менеджеры NoSketchEngine и CWB/CQP, уже требовала наличия VPS-хостинга или выделенного сервера с полным доступом ко всем ресурсам операционной системы GNU/Linux. При этом серьезно возросли требования к объему оперативной (ОЗУ) и дисковой памяти. На сегодняшний день это 16 Гб и 30 Гб соответственно.

10. Дополнительные сервисы и API

Как известно, построение корпуса обычно преследует вполне конкретные цели в области академического исследования языка, но, как показала практика, эти цели могут быть распространены на вполне прикладные задачи. Наличие огромной структурированной информации позволяет авто-

рам Письменного корпуса татарского языка использовать его в том или ином виде для создания или улучшения многих прикладных проектов. Данная тема выходит за рамки текущей статьи, поэтому перечислим лишь некоторые из них:

а) Онлайн система проверки правописания.

б) Генерация различных уникальных статистических данных (частотные списки букв и их комбинаций, слов, лемм, n-грамм).

в) Тезаурус, где размещены векторные представления слов, сгенерированные на базе неглубокой нейронной сети через технологию word2vec.

г) Система синтеза татарской речи, разработанная в Республиканской специальной библиотеке для слепых и слабовидящих.

д) Система распознавания татарской речи Common Voice.

е) Оценка покрываемости для морфологического анализатора Apertium, генерация для включения в словарь списка нераспознанных словоформ, поиск ошибок в наборе правил двухуровневого представления данных и др.

Некоторые из этих проектов в качестве веб-сервиса могут привлекать дополнительных пользователей, что повышает востребованность и популярность проекта в целом.

На сегодняшний день проект Письменного корпуса татарского языка продолжает активно разрабатываться энтузиастами-волонтерами и регулярно используется различными исследователями в сфере татарского языка, несмотря на отсутствие какого-либо государственного финансирования в течение всех девяти лет развития.

Библиография

Захаров В. П., Богданова С. Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и доп. СПб., 2013.

ПКТЯ – Письменный корпус татарского языка. Режим доступа: http://www.corpus.tatar

Сайхунов М. Р., Хусаинов Р. Р., Ибрагимов Т. И. Система сложного морфологического поиска в Письменном корпусе татарского языка // Традиционная культура тюркских народов в изменяющемся мире: Материалы I Междунар. науч. конф. (12–15 апреля 2017 г.). Казань: Изд-во «Ак Буре», 2017. С. 382–385.

Сайхунов М. Р., Хусаинов Р. Р., Ибрагимов Т. И. Эволюция систем поиска в Письменном корпусе татарского языка // Языковые контакты народов Поволжья и Урала: Сб. ст. XI Междунар. симпозиума (Чебоксары, 21–24 мая 2018 г.) / Сост. и отв. ред. А. М. Иванова, Э. В. Фомин. Чебоксары: Изд-во Чуваш. ун-та, 2018. С. 314–319.

Curl. Режим доступа: https://curl.haxx.se/

Heritrix – a web crawler. Режим доступа: https://github.com/internetarchive/heritrix3/wiki

NoSketch Engine. Режим доступа: https://nlp.fi.muni.cz/trac/noske

The IMS Open Corpus Workbench (CWB). Режим доступа: http://cwb.sourceforge.net

Washington J. N., Salimzyanov I. F., Tyers F. M. Finite-state morphological transducers for three Kypchak languages // Proceedings of the 9th Conference on Language Resources and Evaluation. Reykjavik, 2014.

Wget. Режим доступа: https://www.gnu.org/software/wget/

УДК 811.161.1+811.512.142

Сиразитдинов Зиннур Амирович
Институт истории, языка и литературы УФИЦ РАН
г. Уфа, Республика Башкортостан

О ПРИМЕНЕНИИ КОРПУСОВ В ПРАКТИКЕ РУССКО-БАШКИРСКОГО ПЕРЕВОДА

Аннотация: В статье рассматривается применение башкирских лингвистических корпусов в переводческой практике. Проведенное исследование показывает эффективность использования корпусов в русско-башкирском переводе для повышения качества работы, оптимизации переводческой деятельности. В работе отмечается, что даже при наличии словарей у переводчиков нередко возникают сложности, связанные с выбором правильного переводного эквивалента из ряда представленных в словарной статье. Данные национальных языковых корпусов могут использоваться для уточнения значения и закономерностей функционирования лексических единиц языка.

Ключевые слова: национальные языковые корпусы, русско-башкирский перевод, переводческая практика, словарная статья.