М. Р. Сайхунов , Р. Р. Хусаинов, Т. И. Ибрагимов

ЭВОЛЮЦИЯ СИСТЕМ ПОИСКА В ПИСЬМЕННОМ КОРПУСЕ ТАТАРСКОГО ЯЗЫКА

Аннотация. Работа посвящена описанию истории развития систем поиска в Письменном корпусе татарского языка. В первой версии Корпус был способен выполнять поиск по отдельным словоформам, позже постепенно появлялись функции поиска с учетом морфологических категорий, масок, регистра букв. На сегодняшний день внедрена система поиска по n-граммам.

Ключевые слова: корпусная лингвистика, татарский язык, обработка естественного языка, быстрый поиск, морфология, n-граммы.

M. R. Saikhunov, R. R. Husainov, T. I. Ibragimov

SEARCH SYSTEMS EVOLUTION IN THE CORPUS OF WRITTEN TATAR LANGUAGE

Abstract. The work is devoted to the description of search systems development history in the Corpus of Written Tatar language. The first version of the Corpus was able to search by individual word forms, later search functions taking into account morphological categories, masks, letters' register gradually appeared. To date, a search system by n-grams has been implemented.

Keywords: corpus linguistics, the Tatar language, natural language processing, fast search, morphology, n-grams.

1. Письменный корпус татарского языка. Письменный корпус татарского языка функционирует в сети интернет с 2012 г. [Сайхунов, 2012] и включает в себя самые разные тексты на татарском языке общим объемом более 116 миллионов словоупотреблений. Количество предложений в корпусе превышает 10 миллионов.

С момента создания письменный корпус активно развивается. Данный проект не финансируется какими-либо научными фондами или организациями. Все работы над корпусом татарского языка ведутся участниками проекта исключительно в свободное время.

На сегодняшний день корпус также обладает двумя встроенными системами синтеза татарской речи, которые позволяют прослушивать как найденные в результате поиска предложения, так и произвольный текст любого объема. Помимо того, на сайте корпуса размещена система онлайн-проверки орфографии текстов на татарском языке. Авторы размещают различные дополнительные статистические материалы по мере их получения в результате обработки корпуса (как при проведении собственных исследований, так и на основе поступивших внешних предложений), а именно: списки наиболее часто употребляемых словоформ татарского языка; словосочетаний, состоящих из двух-шести элементов; списки частотностей лемм татарского языка, сгруппированных по частям речи; грамматических форм татарского языка; букв и их сочетаний в различных позициях слов; фонем и их сочетаний в пределах слова и ритмической группы.

2. Система поиска Лейпцигской коллекции корпусов. Изначально Письменный корпус татарского языка был спроектирован по аналогии с системой поиска Лейпцигской коллекции корпусов, особенностью которой является работа с контекстами слов. Поиск мог выполняться лишь по словоформе, однако в качестве результата пользователь получал не только набор предложений, но и отсортированные по частоте списки словосочетаний (правый и левый контексты). Система позволяла также просмотреть так называемый семантический контекст, в котором отражается вероятность нахождения различных слов с искомым в рамках одного предложения.

3. Поиск по началу и концу слова. Данный этап являлся временным промежуточным решением для исследования различных новых возможностей, которые было необходимо расширять в корпусе. Идея заинтересовала пользователей, однако обнаружился явный недостаток гибкости системы в плане построения сложных поисковых запросов с учетом грамматических показателей.

4. Создание системы сложного морфологического поиска. В целях расширения функциональных возможностей корпуса татарского языка в 2014 г. нами была произведена морфологическая разметка. Для этого использовалась разработанная международным проектом Apertium [Apertium] система автоматической грамматической аннотации, которая поддерживает большое количество языков (в том числе и татарский).

Основными факторами в пользу выбора системы Apertium являются высокое качество морфологической разметки; наличие универсальной системы тегов для тюркских языков, что подразумевает перспективу выстраивания различных лексических и грамматических связей между корпусами разных тюркских языков; полная открытость исходных кодов и всех наработок (словари, правила).

В связи с появлением в корпусе большого объема новой метаинформации в том же 2014 г. начата работа над новой корпусной поисковой системой, удовлетворяющей следующим критериям:

– поддержка таких параметров поиска, как словоформа, лемма, грамматические (морфологические) теги, маска (шаблон), учет регистра (прописные и строчные буквы), расстояние между словами;

– возможность создания различных комбинаций на базе указанных параметров;
– простой синтаксис запросов, понятный для большинства пользователей;
– высокая скорость поиска.

В ходе работ был учтен опыт различных известных проектов, среди которых TSCorpus [Aksan, 2009], (No)Sketch Engine [Kilgarriff, 2003] и др. [Jurafsky, 2009]. Разрабатываемая нами система в целом получила название «Сложный морфологический поиск». В качестве ядра был создан корпусный поисковый движок «Fastmorph», который успешно выполняет все указанные выше задачи.

На сегодняшний день система сложного морфологического поиска активно используется в составе Письменного корпуса татарского языка. Работы по внедрению дополнительных возможностей продолжаются.

5. Возможности системы сложного морфологического поиска. Рассмотрим несколько примеров возможных запросов в системе «Сложный морфологический поиск». Следует помнить, что фигурные скобки здесь даны только для указания на соответствующие текстовые поля на странице поиска и не используются в реальных запросах.

Для начала произведем поиск по комбинации *«{<adj>} 1-2 {<n><dat>} 1-3 {<v><past>}»*.

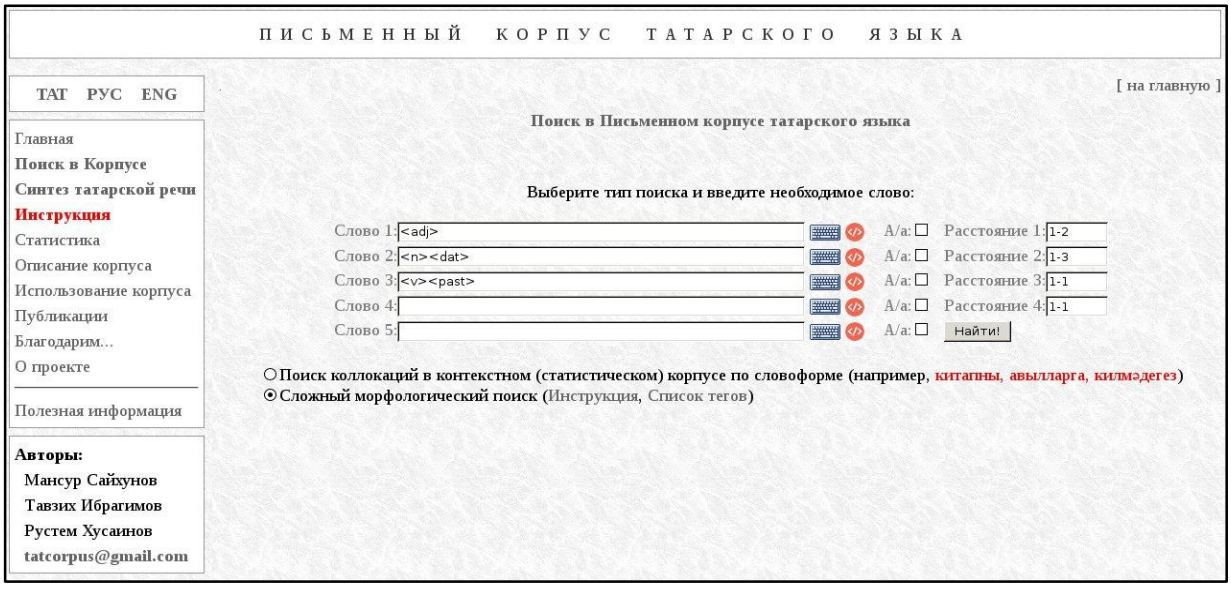

Рис. 1. Образец поискового запроса

Данное выражение означает, что первое слово должно быть прилагательным (<adj>), следующее за ним на расстоянии от одного до двух слов должно быть существительным (<n>) *в дательном падеже* (<dat>), а после него на расстоянии до трех слов должен идти *глагол* (<v>) в форме *-ган* / *-гән* / *-кан* / *-кән* прошедшего времени (<past>).

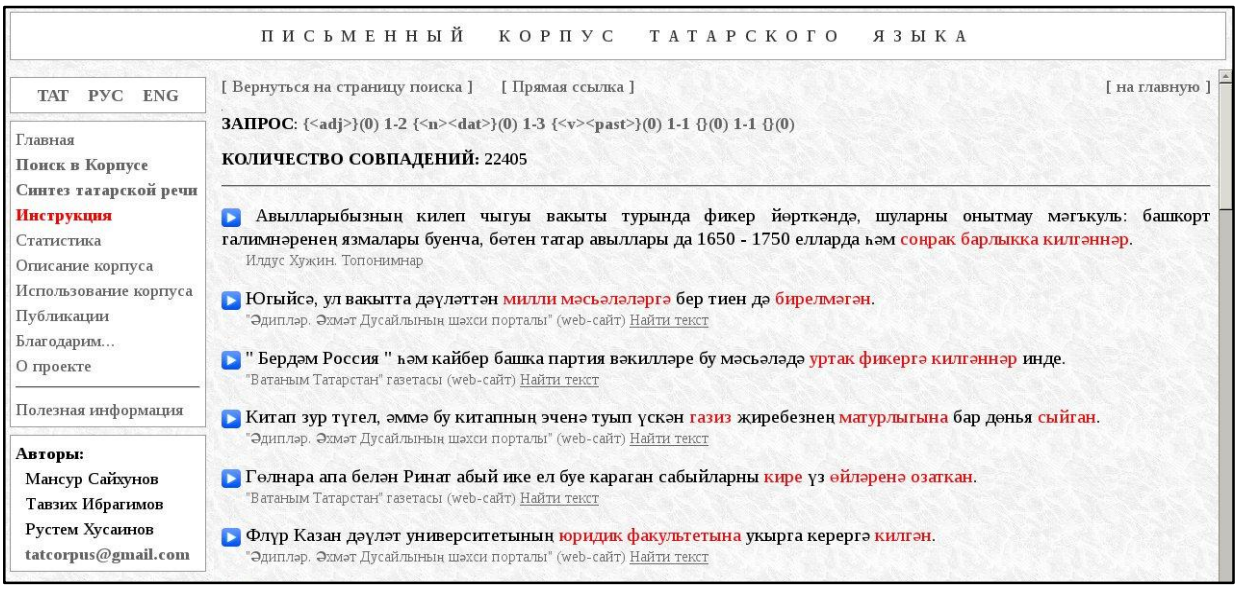

Рис. 2. Результат поискового запроса

В качестве второго примера укажем параметры *«{ил*} 1-1 {белән}»,* которые означают, что первое слово должно начинаться на *ил,* а следующее непосредственно за ним слово должно быть *белән* (послелог в значении предлога ‘с’).

Можно указать среднюю часть слова, используя запрос вида *«*әме*»,* который соответствует словам *керәмен, әмер, үсәме*.

Шаблон поиска по концу слова выглядит как «**рны*». В результате получим предложения со словами *дусларны, атларны, барны, кулларны* и т. п.

Для поиска по началу, средней части и по концу слова можно оформить запрос в виде *«к*әме*н»,* что в итоге приведет к нахождению, например, *керәмен, каләмен, күләменнән, кияүдәмен.*

Знак звездочка «*» совпадает с любым количеством (от нуля до бесконечности) любых символов, а знак вопроса («?») соответствует любому одиночному символу. Например, по образцу «*т?з**» будут найдены такие слова, как *тиз, тозны, түзде, тазарды*, но не *тигез, тугызны, тәрәзә.*

Все перечисленные поисковые параметры (словоформа, лемма, грамматические теги, шаблоны) могут быть комбинированы различными способами. Например, по запросу *«{<prn>} 1-1 {(кеше)} 1-3 {ал*}»* будет произведен поиск всех совпадений, где первое слово является местоимением (<prn>), следующее непосредственно за ним слово является одной из форм леммы *кеше* ‘человек’, а расположенное на расстоянии до трех слов слово, начинающееся на *ал*.

В качестве еще одного примера рассмотрим следующую ситуацию. Допустим, что необходимо найти случаи употребления словоформы *алма* ‘не бери’, однако в результаты поиска попадут также предложения с омонимом *алма* ‘яблоко’. Для того чтобы исключить последние совпадения, можно поставить морфологический тег *«<v>»,* определяющий данное слово как *глагол: «алма<v>».*

В этом случае имеется и другой способ решения данной проблемы. Можно вместе со словоформой *алма* указать соответствующую ей лемму, т. е. *(ал)* ‘брать’. В итоге запрос примет вид «*алма(ал)*». Таким образом, система будет искать только те случаи *алма,* где леммой данной словоформы является *ал,* опуская при этом все результаты, где лемма – *алма* ‘яблоко’.

Технически пользователь может даже оформить запрос в виде *«алма(ал)<v>»* или *«(ал)алма<v>».* Система выполнит разбор данного выражения и будет искать примеры со словоформой *алма,* которая представлена леммой *(ал)* и имеет морфологический тег *«<v>»,* означающий глагол.

Для тех пользователей, которым подобный текстовый способ ввода различных параметров поиска может показаться неудобным, разработан удобный графический режим, где грамматические теги можно выбирать, проставляя галочки и система сама правильно оформит написание леммы, начала, середины и конца слова.

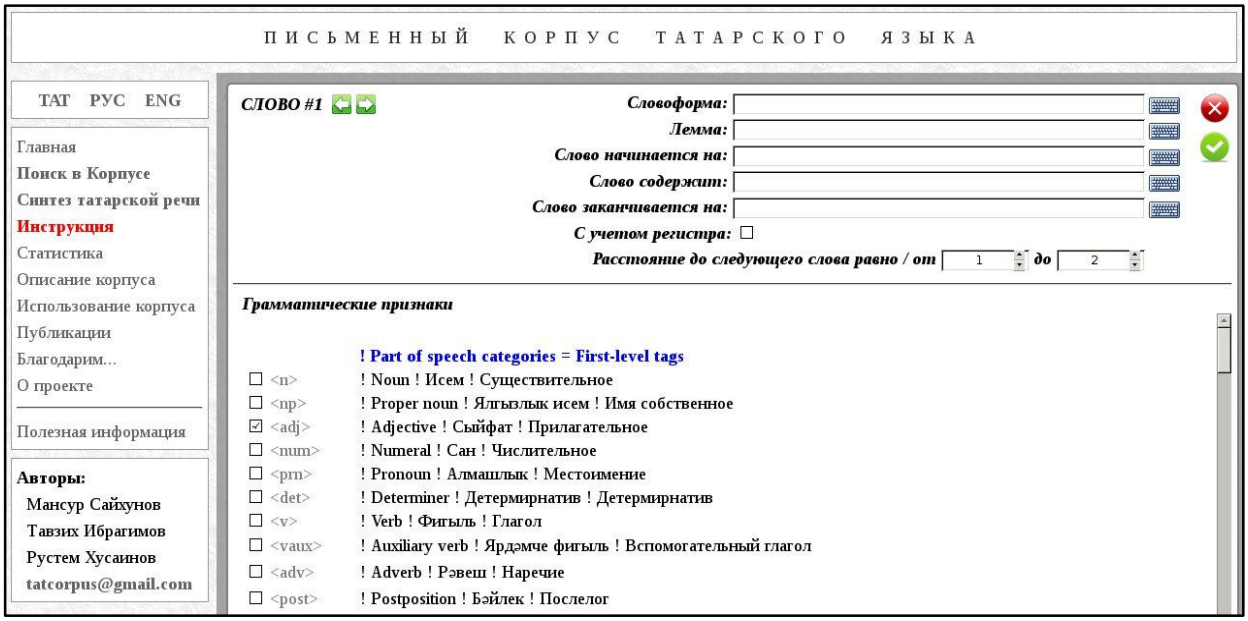

Рис. 3. Графический режим ввода параметров поиска

Руководство пользователя на русском, татарском и английском языках, описывающее все возможности Письменного корпуса татарского языка расположено на сайте в разделе «Инструкция» [Сайхунов, 2015].

6. Техническое описание проекта. Система сложного морфологического поиска устроена следующим образом. Было принято решение написать серверную часть поиска на языке программирования C в связи с тем, что необходимо обеспечить быстрый поиск по таким параметрам, как словоформа,

лемма, морфологические теги, шаблоны, расстояния между словами. Иными словами, данный модуль с рабочим названием Fastmorph изначально задумывался как изолированная по отношению к веб-серверу система. Fastmorph на этапе инициализации загружает все необходимые данные корпуса из СУБД MySQL и компактно размещает их в виде массивов в оперативной памяти компьютера, что позволяет избежать потери времени при операциях обращения к жесткому диску, делает архитектуру приложения максимально простой и гибкой для дальнейшего расширения функциональности или адаптации под другие проекты.

Реализованные на языке PHP дополнительные программы производят преобразование относительно свободного запроса пользователя в строго типизированный набор данных и передают системе Fastmorph. Fastmorph выполняет поиск в своей индексированной базе и возвращает список найденных предложений с метаинформацией (список источников, выделение в предложениях искомых элементов специальными тегами, указание лемм и набора тегов найденных элементов, общее количество найденных примеров и др.) обратно PHP модулю, а тот, в свою очередь, – пользователю.

Заложенная в основу алгоритма настоящая идея позволяет эффективно утилизировать как кэш процессора, так и вычислительные мощности мультиядерных систем. В итоге удалось добиться большой скорости выполнения поиска в 0,2–2 секунды в зависимости от параметров поиска даже на скромном оборудовании. При этом не используются такие ухищрения, как отображение примерного количества совпадений и отложенный (фоновый) поиск.

22 ноября 2016 г. исходный код корпусного поискового движка Fastmorph был открыт [Fastmorph] под лицензией GNU General Public License v3 [GNU GPL v3], что позволяет всем желающим использовать наши наработки в своих проектах.

7. N-граммы по шаблону / маске. В целях минимизации потери времени пользователей на просмотр результатов поиска была разработана система поиска по n-граммам на основе масок, т. е. подстановочных знаков «*» и «?». N-граммы представляют собой последовательности n-слов в тексте. На основе данных корпуса были сгенерированы 2, 3, 4, 5 и 6-граммы.

8. N-граммы в системе Fastngrams. Система поиска по n-граммам в корпусе полностью продемонстрировала свою эффективность в плане структуризации найденной в корпусе информации. Однако в текущем ее виде она имела ряд существенных недостатков: большое время поиска, в некоторых случаях до одной минуты и более; невозможность учета в поиске грамматических тегов, лемм, регистра букв. В связи с этим было принято решение о реализации возможности поиска по n-граммам по аналогии с системой Fastmorph. Данная работа заняла около года, и в ноябре 2017 г. была представлена первая версия системы Fastngrams, которая позволила достичь поставленных задач: скорость поиска не превышает нескольких секунд; появилась возможность комбинирования различных поисковых параметров.

Для демонстрации возможностей системы Fastngrams произведем поиск по комбинации «*{<prn>}(0) {(кеше)}(0) {*}(0) {ал*}(0)*» (рис. 4). Первым словом в этой 4-грамме должно быть любое местоимение, леммой второго слова является *кеше*. Третье слово представлено символом «*», что подразумевает любое слово, в том числе и знак препинания. Крайним словом здесь является выражение *ал**, т. е. любое слово, начинающееся на *ал.*

Рис. 4. Образец запроса в поиске по n-граммам

Результатом поиска является информация о количестве найденных совпадений и список из не более чем 1000 наиболее частотных n-грамм, отсортированный по частотности в порядке убывания (рис. 5). Необходимо отметить, что сочетания слов, различающиеся регистром букв и / или грамматическими тегами, представлены отдельными n-граммами.

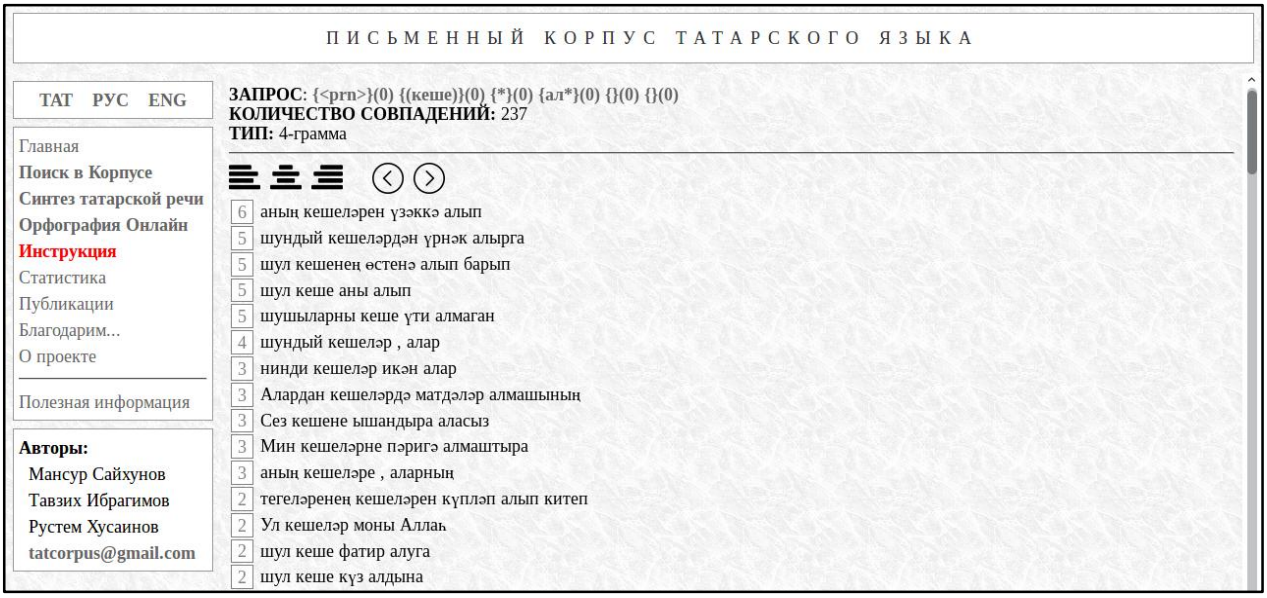

Рис. 5. Результат поиска по n-граммам

Для удобства просмотра имеются инструменты выравнивания n-грамм по левому краю, середине и правому краю, а также выравнивания отдельных слов (рис. 6).

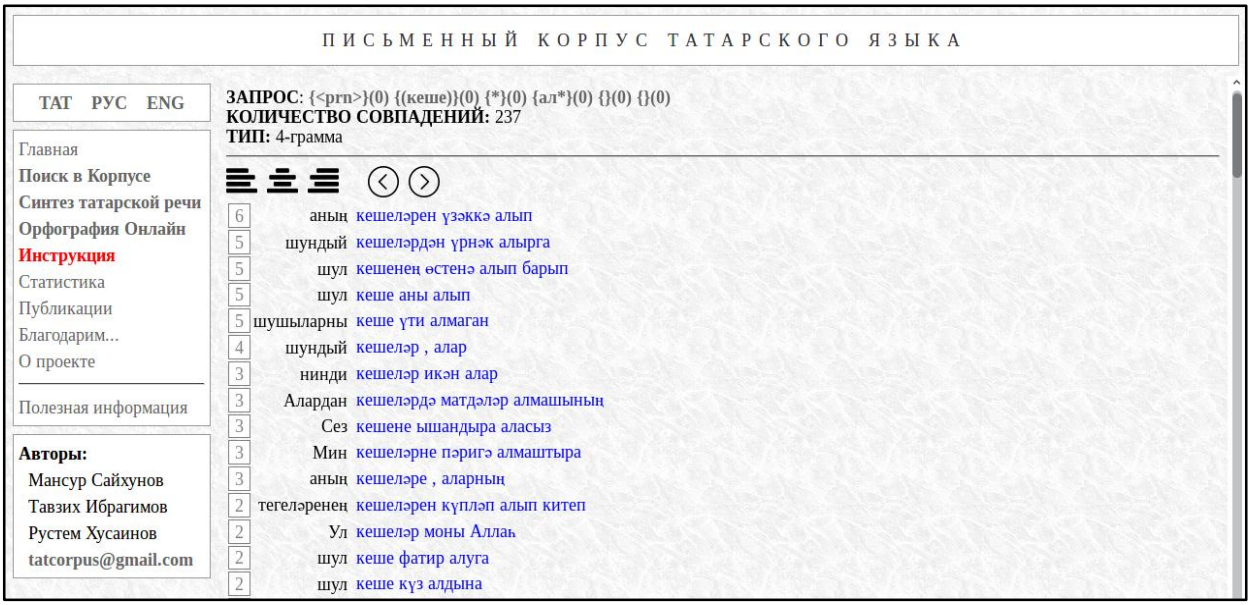

Рис. 6. Выравнивание по второму слову в списке n-грамм

При нажатии на ту или иную n-грамму открывается список предложений, где употребляется данная n-грамма.

9. Планы:

– внедрение системы корпусных запросов CQL [Corpus Query Language], которая на сегодняшний день является де-факто стандартом в корпусной лингвистике;

– добавление возможности включения специализированных подкорпусов;

– расширение функционала фильтра поиска: по авторам, жанрам, произведениям, диапазонам дат и др.;

– дальнейшее расширение сгенерированных статистических материалов;

– разработка вспомогательного набора утилит для облегчения использования систем Fastmorph и Fastngrams в корпусах других языков.

Библиография

Сайхунов М. Р., Ибрагимов Т. И., Хусаинов Р. Р. Письменный корпус татарского языка [Электрон/ ресурс]. РЕЖИМ ДОСТУПА: http://corpus.tatar (дата обращения: 16.12.2017).

Сайхунов М. Р., Ибрагимов Т. И., Луутонен Й. Письменный корпус татарского языка: руководство пользователя [Электрон. ресурс]. РЕЖИМ ДОСТУПА: http://corpus.tatar/manual.htm (дата обращения: 21.12.2017).

Aksan Y., Aksan M. Building a national corpus of Turkish: Design and implementation // Working Papers in Corpus-based Linguistics and Language Education. Tokyo: TUFS, 2009. № 3. P. 299–310.

Apertium – Открытая платформа машинного перевода [Электрон/ ресурс]. Режим доступа: http://wiki.apertium.org/wiki/Publications (дата обращения: 10.01.2018).

Corpus Query Language (CQL) Tutorial [Электрон. ресурс]. Режим доступа: http://cwb.sourceforge.net/files/CQP_Tutorial (дата обращения: 21.12.2017).

Fastmorph – Fast corpus search engine originally made for the Corpus of Written Tatar language [Электрон/ ресурс]. Режим доступа: https://github.com/mansayk/fastmorph (дата обращения: 11.01.2018).

GNU GPL v3 – GNU General Public License v3 [Электрон. ресурс]. Режим доступа: https://www.gnu.org/licenses/gpl-3.0.en.html (дата обращения: 21.12.2017).

Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall, 2009. 1024 p.

Kilgarriff A. Linguistic search engine // In Kiril Simov, editor, Shallow Processing of Large Corpora: Workshop held in association with Corpus Linguistics. Lancaster, 2003. P. 53–58.