

Комитет науки Министерства образования и науки  
Республики Казахстан  
РГП «Институт информационных и вычислительных технологий»  
КН МОН РК



25 лет  
Независимости  
Республики Казахстан



25 лет  
Институту  
информационных и  
вычислительных  
технологий

### МАТЕРИАЛЫ

Международной научной конференции  
«Информатика и прикладная математика»  
(«Computer science and Applied Mathematics»),  
посвященной 25-летию Независимости Республики Казахстан и  
25-летию Института информационных и  
вычислительных технологий

Часть I

г. Алматы, 21-24 сентября 2016 года

Алматы  
2016

	<i>Жанысбекова Г.А., Сембиев О.З., Сулейменова Л.А., Ашим Ж.К., Тимабай Ж.Н.</i>	Локализация на карте ГИС территорий военно-политических процессов Казахского ханства с использованием IT-технологии	111
8	<i>Исмаилов Б.И., Мамбеталиева С.М.</i>	Снижение субъективности весовых коэффициентов при оценке эффективности работы университетов по множеству критериев	125
18	<i>Костылев А.О., Скопина Л.В.</i>	Принятие инвестиционных решений в нефтегазовой отрасли в условиях неопределенности	134
21	<i>Култешов Б.Ш.</i>	Почти счетная категоричность в упорядоченных структурах	141
27	<i>Купчишин А.И., Купчишин А.А., Шмыгалева Т.А., Шмыгалев Е.В.</i>	Основы каскадно-вероятностного метода	151
36	<i>Мансурова М.Е., Койбагаров К.Ч., Барахнин В.Б., Солтангельдинова М., Бердибеков С.</i>	Применение морфологического анализатора казахского языка для извлечения фактов из фактографических систем	156
37	<i>Мурзабеков З.Н., Мурзабеков А.З.</i>	Синтез пропорционально-интегральных регуляторов для систем автоматического управления при наличии ограниченных управляющих воздействий	166
47	<i>Мухарлямов Р.Г.</i>	Построение уравнений динамики систем с линейными дифференциальными программными связями	174
54	<i>Плесневич Г.С., Карабеков Б.С., Нгуен Тхи Минь Ву</i>	Спецификация онтологий для потоков задач	182
63	<i>Рысбайулы Б., Карашибаева Ж.О.</i>	Задача продолжения для процесса переноса тепла и влаги в многослойной области	193
66	<i>Рысбайулы Б., Юничева Н.Р.</i>	Итеративно-интервальный подход к расчету тепловых параметров в метастабильных горных системах при неточных исходных данных	201
76	<i>Глеубергенов М.И., Ажымбаев Д.Т.</i>	О построении стохастических дифференциальных уравнений по заданному интегральному многообразию	210
85	<i>Хисамиев Н.Г., Конырханова А.А.</i>	Вычислимые ретракты разрешимых групп	218
93			
05			

## ПРИМЕНЕНИЕ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА КАЗАХСКОГО ЯЗЫКА ДЛЯ ИЗВЛЕЧЕНИЯ ФАКТОВ И ФАКТОГРАФИЧЕСКИХ СИСТЕМ

Мансурова М.Е.<sup>1</sup>, Койбагаров К.Ч.<sup>2</sup>, Барахнин В.Б.<sup>3,4</sup>,  
Солтангельдинова М.<sup>1</sup>, Бердибеков С.<sup>1</sup>

<sup>1</sup>Казахский национальный университет имени аль-Фараби, Казахстан

<sup>2</sup>Институт информационных и вычислительных технологий МОН РК,  
Казахстан,

<sup>3</sup>Институт вычислительных технологий СО РАН, Россия,

<sup>4</sup>Новосибирский государственный университет, Россия

### Аннотация

В данной работе описывается применение лингвистического анализатора для казахского языка для автоматизированного наполнения онтологии фактографической поисковой системы. Для осуществления морфологической разметки используются лексический и морфологический модули анализатора. Задачей лексического модуля является определение границ предложений, выделение слов, идентификаторов и пунктуационных маркеров. Морфологический модуль выполняет поиск слов в словаре казахского языка и определяет их морфологические метры. Осуществляемая с помощью разработанного анализатора морфологическая разметка является этапом автоматизированного наполнения онтологии фактографической поисковой системы.

### 1 Введение

В тюркологии существует большое количество исследований по автоматизированной морфологической разметке текстов для родственных казахскому языков, основанных на разных концептуальных подходах [1, 2, 3, 4, 5]. Анализ открытых публикаций в области технологий морфологического анализа словоформ казахского языка показывает, что данное направление исследований представляет большой теоретический и практический интерес.

С 1970-2000 года публикации в области морфологии казахского языка носят в основном теоретический характер. С 2006 годов появились публикации в зарубежных журналах ([6, 7, 8, 9]). В работах российских исследователей описаны принципы разбиения существительных казахского языка на флективные классы и построение алгоритма на этом разбиении алгоритма синтеза словоформ [10, 11]. Среди казахстанских ученых особо следует отметить работу Шарипбаева А.А. ([12]), посвященную проектированию интеллектуального морфологического анализатора, основанного на семантических сетях. Целый ряд работ исследовательской группы из Назарбаевского Университета посвящен методу морфологического анализа и устранения неоднозначности для казахского языка, который учитывает как флективную, так и деривационную морфологию [13, 14, 15]. Все перечисленные работы затрагивают различные аспекты области морфологии и синтаксиса казахского языка, при этом основное внимание уделяется методологии построения морфологического анализатора. В связи с этим актуальными являются вопросы проектирования, программной реализации и предоставления в открытом доступе морфологического анализатора для казахского языка. Среди методов морфологического анализа, используемых в лингвистических программах, можно выделить методы с декларативной и с процедурной ориентацией. Данная работа

будет продолжением исследований авторов в этой области, в основе которых лежит описанный подход, который использует преимущества как декларативного, так и процедурного методов [16, 17, 18].

Дальнейшее изложение работы выглядит таким образом. В разделе 2 приводится описание разработки анализатора для автоматизации морфологической разметки текста на казахском языке. Раздел 3 содержит сведения о технологии автоматизированного пополнения онтологии фактографической поисковой системы. В разделе 4 описываются результаты вычислительных экспериментов. В разделе 5 формулируются выводы и приводится план будущих исследований.

## 1 Разработка анализатора для автоматизации морфологической разметки текста на казахском языке

### 1.1 Особенности казахской морфологии

Казахский язык относится к классу агглютинативных языков и вместе с узбекским, киргизским, башкирским, татарским, азербайджанским, турецким и другими языками образует тюркскую семью. Для агглютинативных языков характерно последовательное присоединение различных формообразующих суффиксов или окончаний, имеющих грамматическое значение, к неизменяемому корню или основе, являющихся носителями лексического значения.

Порядок добавления аффиксов строго определен. Например, для имен существительных к основе слова вначале добавляется суффикс и далее окончание множественного числа, затем притяжательное окончание, далее следует падежное окончание и окончание формы спряжения (добавляется только к одушевленным существительным) [19, 20].

Для казахского языка существует закон сингармонизма: гармония гласных и согласных звуков аффикса со звуками корня. Гармонируют гласные по принципу твердости и согласные – конечный звук корня и первый звук аффикса. Помимо трех основных правил сингармонизма, необходимо учитывать следующие правила исключения:

1. Правило удаления глухой согласной в прибавляемом аффиксе, если в окончании присутствуют две глухие согласные. Например: журналист+тер -> журналистер.

2. Закон сингармонизма не соблюдается в следующих случаях следующих аффиксов:

а) для аффиксов *мен, пен, бен*: қаламмен; нікі, дікі, тікі: баланікі;

б) для заимствованных слов с окончаниями: *рк, нк, кс, кт*: - пукните

3. Правило выпадения гласной *і, ы* в корне слова при добавлении притяжательного аффикса *і, ы*. Например: Әріп - әрпі, қауіп - қауіпі, қойын - қойыны, Ерін - ерні, құлқы - құлқы.

### 1.2 Структура лингвистического анализатора

На рис. 2 представлен разрабатываемый нами лингвистический анализатор, который состоит из четырех анализаторов (лексический, морфологический, синтаксический и семантический) (рис 1). Анализаторы располагаются последовательно друг за другом, выходной поток одного анализатора служит входом для следующего анализатора. Задача лексического анализатора является определение границ предложений, выделение идентификаторов и пунктуационных маркеров. Морфологический анализатор выполняет поиск слов в словаре (словарь представляет собой отдельную базу данных) и определяет их морфологические параметры (например, часть речи, число, падеж и т.д.). Синтаксический анализатор выполняет построение синтаксического графа

предложения. В данной работе мы используем два анализатора – лексический и морфологический анализатор.

Лексический анализатор – это программа начального анализа естественного языка, представленного в виде цепочки Unicode символов. Выходная информация необходима для дальнейшей обработки морфологическим и синтаксическим анализаторами. В задачу лексического анализатора входят:

1. Разделение входного текста на слова, числа, разделители и т.д.
2. Выделение устойчивых оборотов, не имеющих словоизменительных вариантов;
3. Выделение имен собственных, выделение ФИО (фамилия, имя, отчество), когда имя и отчество написаны инициалами;
4. Выделение электронных адресов и имен файлов;
5. Выделение предложений из входного текста;

Процедура выделения из текста слов, чисел и знаков препинания совершенно очевидна. После считывания очередного абзаца текста графематический анализатор разбивает токены и приписывает им соответствующие графематические характеристики. На этом этапе выделение токенов производится по пробелам и знакам препинания. Однако наибольшую сложность составляет определение начала и конца предложения. Лексический анализатор содержит эвристический механизм определения границ предложения и результатами лексического анализа является не только массив лексем, но и указатели на начало и конец текущего предложения в тексте.

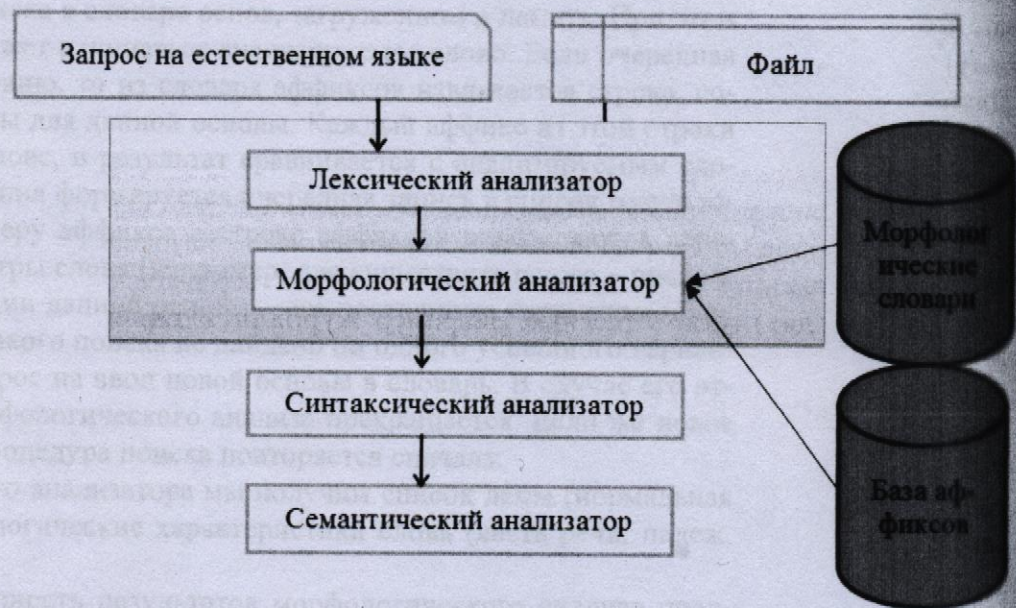


Рис. 1. Состав лингвистического анализатора

Найти конец предложения в тексте тоже не так просто, как может показаться. Восклицательный или вопросительный знак наверняка означают конец предложения, но вот точка может стоять и после сокращения, и в середине десятичной дроби. Следует также учитывать и сложные единицы измерения (кв.м, км/час), интернет-адреса (<http://yandex.ru>), записанные цифрами порядковые числительные (1917-ж.), выделенные фамилий (Касымов К.С.). Рассмотрим следующий фрагмент:

1917 ж. 21-26 шілдеде Покровский С.И. Орынборда болған «Бүкілқазақтық»

Здесь точка встречается четырежды, и только в четвертом случае она означает координату предложения. Поэтому в анализатор введены специальные блоки проверки. В частности, простейшая проверка: лексема, стоящая непосредственно перед точкой, должна содержать хотя бы одну гласную. Такой проверкой удастся отобрать многие сокращенные употребляемые с точкой на конце:

Ж., Т.Г.К., ТГ., Ф.-М.Г.Д., К.

Разумеется, такая проверка не гарантирует правильность результата. С одной стороны, сокращение или число с точкой могут действительно стоять в конце предложения. Тем не менее, опыт работы с документами подтверждает эффективность такой проверки. Только после анализа точек, не являющихся маркером предложения, выполняется разбивка абзаца на предложения, и весь дальнейший анализ ведется уже только в пределах одного предложения.

### 1.3 Описание алгоритма морфологического анализатора

Работает морфологический анализатор следующим образом. На его вход поступают слова, знаки препинания и чисел, выделенных из входного текста на этапе лексического анализа, с лексическими характеристиками [21]. Для каждого слова анализатор выполняет процедуру поиска в словаре основ, загруженном в память. При этом проверяются все основы, с которых может начинаться анализируемое слово. Если очередная основа удовлетворяет этому условию, то из словаря аффиксов извлекается строка, содержащая все возможные аффиксы для данной основы. Каждый аффикс из этой строки поочередно присоединяется к основе, и результат сравнивается с анализируемым словом. В случае их точного совпадения формируется очередная запись в список результатов поиска: по порядковому номеру аффикса в строке аффиксов определяются параметры морфологические параметры слова (например, для существительного – число и род), а по словарной информации данной основы – его постоянные параметры (сущ., прил...). Если в результате такого поиска не найдено ни одного успешного варианта, то пользователю выдается запрос на ввод новой основы в словарь. В случае его отклонения это сделать выполнение морфологического анализа прекращается. Если же новое слово введено в словарь, то вся процедура поиска повторяется сначала.

На выходе морфологического анализатора мы получим список лемм (нормальная форма слова) + аффикс + морфологические характеристики слова (часть речи, падеж, склонение).

Таким образом, вся совокупность результатов морфологического анализа представляется в виде динамического массива. Число его элементов равно числу лексем в предложении. Элементами массива являются другие массивы, каждый из которых хранит все возможные интерпретации своей лексемы омонимы. В качестве исходных лексических материалов используются словарь основ слов, словарь географических названий, словарь фамилий, словарь аффиксальных соединений.

Разработанный морфологический анализатор представлен в открытом доступе: <http://imf.kaznu.kz/>. Наличие морфологической разметки текстов корпуса расширяет возможности поиска по корпусу: в качестве поискового запроса пользователь будет иметь возможность задать любое сочетание признаков слова, имеющихся в системе. Разработанный анализатор позволяет решать различные задачи по обработке текстов на казахском языке. В следующем разделе описывается применение морфологического анализатора для автоматизированного наполнения онтологии фактографической поисковой системы.

### **3 Алгоритм автоматизированного наполнения онтологии фактографической поисковой системы**

В данном разделе описан алгоритм, позволяющий извлечь ключевые слова/словосочетания из корпуса текстов однородной тематики. Извлеченные ключевые слова в дальнейшем используются в качестве возможных значений атрибутов сущностей, описываемых в создаваемой онтологии предметной области, предназначенной для организации фактографического поиска. В качестве входных данных используется текст на казахском языке, предварительно размеченный с помощью морфологического анализатора, описанного в разделе 2. Для извлечения семантически связанных слов/словосочетаний в алгоритме применяется метод случайных блужданий по набору этих словосочетаний с целью отнесения конкретного словосочетания к определенному атрибуту описываемой в тексте сущности применяется обученная нейронная сеть со скрытым слоем.

#### **3.1 Автоматизированное извлечение фактов в системах фактографического поиска**

Важным этапом процесса функционирования фактографических систем является извлечение фактов, содержащихся в текстах документов. Как показано в [22], под фактом понимается совокупность связей между сущностями, содержащаяся в тексте заданного документа. Практическая реализация такого подхода опирается на представление множества сущностей [23], отличительные особенности которой заключаются в том, что, во-первых, в ней всё трактуется как объекты, во-вторых, все связи в этой модели бинарные, причем связи между объектами также рассматриваются как объектные данные, в свою очередь, с объектами – атрибутами связей. Таким образом, на практике в качестве атомарного факта можно рассматривать входящую в текст и в метаданных документа характеристику сущности, описываемую в онтологии информационной системы. Такая характеристика представляется как единичное значение данных, в дальнейшей работе под онтологией понимается модель с произвольным набором отношений.

Наиболее сложной проблемой, возникающей при создании фактографических систем, является разработка методик автоматизированного извлечения фактов из документов на естественном языке. Для достижения поставленной цели необходимо решить следующие задачи:

- 1) разметить текст, определяя, к какой части речи относится каждая лексема, входящая в текст;
- 2) извлечь из размеченного текста наборы семантически связанных слов/словосочетаний;
- 3) создать нейросеть, обученную извлечению лингвистических конструкций, в которые входят возможные значения атрибутов сущностей, описываемых в онтологии предметной области;
- 4) используя коллекцию лингвистических конструкций и наборы семантически связанных слов, наполнить онтологию предметной области.

Особенностью данного алгоритма является возможность использования результатов его применения для автоматизированного добавления вики-разметки в структурированные документы, предназначенные для помещения в Википедию.

#### **3.2 Создание и обучение нейронной сети**

Для выполнения этого этапа в качестве входных данных выбирается корпус документов с википодобной разметкой. Данный корпус используется для обучения нейронной сети, извлекающей лингвистические конструкции, в которые входят возможные значения атрибутов сущностей, описываемых в онтологии предметной области.

атрибутов сущностей, описываемых в онтологии предметной области. Мы остановились на выборе нейронной сети со скрытым слоем [24].

Известные входные данные – это лексические конструкции, например, для биографии: “родился  $x$ ”, “появился на свет  $x$ ” и т.п., где  $x$  – искомая характеристика (дата рождения). Процесс обучения сети осуществляется посредством предъявления каждого набора данных и последующего распространения ошибки. Стоит отметить, в условиях данной задачи весовые коэффициенты являются булевозначными, равно как и вектор ошибки. В рассматриваемой задаче при распознавании синонимов сначала выбирается синоним первой характеристики, синоним второй и так далее, затем процесс повторяется. В итоге обученная сеть способна извлекать лингвистические конструкции, в которые входят возможные значения атрибутов сущностей, описываемых в онтологии предметной области. Отметим, что в текущей реализации скрытый слой реализован в виде коллекции в MongoDB.

#### 1.3 Извлечение наборов семантически связанных слов

В данной работе на основе морфологического анализа текста с помощью метода случайных блужданий извлекаются семантически связанные слова. Предположим, что имеется текст, включающий в себя  $n$  документов:  $D = (d_1, d_2, \dots, d_n)$ .

В данном случае, в качестве документа при анализе текста целесообразно рассмотреть отдельно взятое предложение. Обозначим через:  $W = (w_1, w_2, \dots, w_n)$  слова в каждом слове документа нам известны его морфологические характеристики (род, число, падеж и т.д.). Кроме того, для каждого документа мы можем задать вес слова в данном документе (TF-IDF).

Анализируемый текст можно представить в виде гиперграфа  $H = (V, E)$ , в котором вершины  $v \in V$  – это слова текста, гиперребра  $e \in E$  – это документы нашего текста. Семантическую связь между вершинами (словами) нашего гиперграфа будем вычислять на основе модели случайных блужданий (random walk model) между вершинами графа. Предполагается, что случайный переход от вершины к вершине осуществляется в два этапа (на первом выбирается одно из гиперребер, смежных с нашей вершиной, на втором этапе случайным образом выбирается вершина данного гиперребра). Таким образом, процесс случайных блужданий будет представлять собой дискретную цепь Маркова с конечным множеством состояний. Как известно, основным инструментом такого процесса является матрица вероятностей переходов между состояниями цепи. Таким образом, на выходе получаем набор семантически связанных слов, на основе которого будет реализован алгоритм фактографического поиска.

#### 1.4 Наполнение онтологии предметной области

Работа алгоритма была проверена на биографиях учёных, так как биографии при определённой структуре имеют сходную онтологию. В качестве обучающей выборки были взяты статьи из Википедии. В качестве примера полного цикла работы алгоритма была взята биография Фаризы Оңғарсынқызы. На первом этапе текст биографии разбит на части речи. Затем с помощью метода случайных блужданий выделяются связанные слова и словосочетания.

#### 1.5 Вычислительные эксперименты

Данный алгоритм был опробован на биографиях выдающихся личностей. Для обучения нейронной сети со скрытым слоем было использовано 100 статей из Википедии и сайта Жарар [25] и извлеченные из соответствующих биографий наборы



ключевых слов. В качестве дескрипторов для извлечения онтологии были выделены следующие характеристики: имя, род деятельности, дата рождения, дата смерти.

На первом этапе, текст биографии размечается по частям речи с помощью морфологического анализатора. При разметке текст разделяется по предложениям. Для каждого слова, входящего в предложение, по базе ищутся все основы, от которых может начинаться данное слово. И таким образом, идет анализ каждого предложения, определяются все характеристики и на выходе мы получаем список лемм + их морфологические характеристики.

Ниже приведена часть размеченного текста:

Предложение: *Фариза Оңғарсынқызы Оңғарсынова - қазақ ақыны, жазушысы, журналист. 1939 жылы 5 желтоқсанда Гурьев (қазіргі Атырау) облысы, Новобогат ауданына қарасты Манап ауылында туған.*

Разметка:

Фариза/фариза|| [fio]  
қазақ/қазақ|| [зт]  
ақыны/ақың|ы| [ЗТ|ы/ТЖ-3+ØАС] // сущ + притяж.оконч  
./,|| [Comma]  
халық/халық|| [зт]  
жазушысы/жазушы|сы| [ЗТ|сы/ТЖ-3+ØАС] сущ + суффикс + притяж.оконч  
./,|| [Comma]  
журналист/журналист|| [зт]  
./,|| [Comma]  
..  
1939/1939|| [NUMCHAR]  
<omonim>  
жылы/жылы|| [сн]  
жылы/жылы|| [ет]  
жылы/жыл|ы| [ЗТ|ы/ТЖ-3+ØАС]  
</omonim>  
5/5|| [NUMCHAR]  
желтоқсанда/желтоқсан|да| [ЗТ|да/ЖС]  
Гурьев/гурьев|| [fio]  
(/|| [LBracet]  
қазіргі/қазіргі|| [сн]  
<omonim>  
Атырау/атырау|| [зт]  
Атырау/атырау|| [geo]  
</omonim>  
)|| [RBracet]  
облысы/облыс|ы| [ЗТ|ы/ТЖ-3+ØАС]  
./,|| [Comma]  
Новобогат/<Unknown>  
ауданына/аудан|ына| [ЗТ|ы/ТЖ-3+на/БС]  
<omonim>  
қарасты/қарасты|| [үс]  
қарасты/қарас|ты| [ЗТ|ты/ТС]  
қарасты/қарас|ты| [ЕТ|ты/ЖӨШ+АР+ЖЖ-3]

manash| [fio]  
manash| [3T|ы/ГЖ-3+нда/ЖС]  
manash| [3T]  
manash| [ES|ған/БС]  
manash| [ET|ған/E+AP+ЖЖ-3]  
manash|

с помощью метода случайных блужданий выделяются ключевые  
слова, например:

*фариза оңгарсынқызы оңгарсынова, ақын, жазушы, 1939 жыл 5 желтоқсан,  
Манаш ауылы и т.д.*

на последнем этапе, нейронная сеть размещает данные по дескрипторам.

*name: "фариза оңгарсынқызы оңгарсынова"*  
*profession: "ақын", "жазушы"*  
*date of birth: "1939 жыл 5 желтоқсан"*  
*date of death: "2014 жылдың 23 қаңтар"*

#### Научное значение

В данной работе описывается применение морфологического анализатора для автоматического наполнения онтологии фактографической поисковой системы. Данная технология основана на применении морфологического анализатора для анализа текста с последующим использованием метода случайных блужданий для извлечения семантически связанных ключевых слов (словосочетаний). К набору этих словосочетаний с целью отнесения конкретного словосочетания к определенному атрибуту, заданной в тексте сущности применяется обученная нейронная сеть со скрытым слоем. В дальнейшем планируется продолжить исследования по разработке модульного морфологического и семантического анализаторов, что расширит возможности для наполнения онтологии и функционирования фактографической поисковой системы.

Исследование выполнено при финансовой поддержке МОН РК в рамках научного проекта № 3033 «Разработка интеллектуальной высокопроизводительной информационно-поисковой системы обработки слабоструктурированных данных». В разработке морфологического анализатора принимали участие сотрудники Института информационных и вычислительных технологий МОН РК и сотрудники Института информатики имени Ахмета Байтурсынова МОН РК.

#### Литература:

- Махмудов М. Системы автоматической переработки тюркского текста на лексико-морфологическом уровне. – Баку: Элм, 1991. – 114 с.  
Миналкин В.В. Моделирование орфографии якутского языка и разработка компьютерных программ для проверки правописания якутских текстов в среде Windows; Автореферат диссертации на соискание ученой степени кандидата технических наук. Якутск, 2005.  
Сидыков Т. Проблемы моделирования тюркской морфологии: аспект порождения именной словоформы. – Изд-во Илим, 1987. – 119 с.

4. Сиразитдинов З.А. Моделирование грамматики башкирского языка. Словоизменительная система. – Уфа, 2006. – 160 с.
5. Сиразитдинов З.А. О моделировании словоизменительной системы агглютивных языков парными сочетаниями (на примере башкирского языка)/Актуальные проблемы современного монголоведения и алтаистики. Материалы Международной научной конференции. Элиста, 2014. С. 139-143.
6. Jonathan North Washington. A Novel Approach to Delineating Kazakh's Five Present Tenses: Lexical Aspect. (2006) <http://jnw.name/papers/2006wi-kzpres.pdf> (дата просмотра: 01.06.2016)
7. Altenbek G. and WANG Xiao-long. Kazakh segmentation system of inflectional fixes. // In Joint Conference on Chinese Language Processing, CIPS- SIGHAN, – p. 183-190. 2010.
8. Zafer H.R., Tilki B., Kurt A., Kara M. Two-level description of Kazakh morphology // In: Proceedings of the 1st International Conference on Foreign Language teaching and Applied Linguistics, FLTAL 2011, Sarajevo (May 2011).
9. Kessikbayeva G., Cicekli I. A Rule Based Morphological Analyzer and a Morphological Disambiguator for Kazakh Language. Linguistics and Literature Studies, 4(1), 96 – 105. 2016.
10. Барахнин В.Б., Лукпанова Л.Х., Соловьев А.А. Алгоритм построения словоформ с использованием флексивных классов для систем морфологического анализа казахского языка // Вестник Новосибирского государственного университета. Серия Информационные технологии. – 2014. – Т. 12, № 2. – С. 25-32.
11. Барахнин В.Б., Лукпанова Л.Х., Соловьев А.А. Алгоритм синтеза словоформ казахского языка с использованием флексивных классов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: тр. 16 Всерос. конф. (RCDL'2014) Дубна, 13-16 окт. 2014 г. – Дубна: ОИЯИ, 2014. – С. 108-112.
12. Шарипбаев А.А., Бекманова Г.Т., Ергеш Б.Ж., Бурибаева А.К., Карабаева М.Х. Интеллектуальный морфологический анализатор, основанный на семантических сетях // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2012): материалы II Международной научно-технической конференции Минск: БГУИР, 2012. – С. 397-400.
13. Makhambetov O., Makazhanov A., Yessenbayev Zh., Matkarimov B., Sabyrgaliyev I., Sharafudinov A. Assembling the Kazakh Language Corpus. // Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). 2013 – P. 1022-1031.
14. Makhambetov O., Makazhanov A., Yessenbayev Zh., Sabyrgaliyev I., Sharafudinov A. Towards a data-driven morphological analysis of Kazakh language // J. Türkiye bilimsel vakfi bilgisayar bilimleri ve mühendisliği dergisi. – Vol. 7, issue 1. – 2014.
15. Makhambetov O., Makazhanov A., Sabyrgaliyev, Yessenbayev Zh. Data-Driven Morphological Analysis and Disambiguation for Kazakh // International Conference on Intelligent Text Processing and Computational Linguistics Springer International Publishing – P. 151-163
16. Койбагаров К.Ч., Мусабаев Р.Р., Калимолдаев М.Н. Разработка лингвистического процессора текстов на казахском языке. / Проблемы информатики. 2014. – № 1. С. 64-72.
17. Койбагаров К.Ч., Амиргалиев Е.Н., Мусабаев Т.Р. Программная реализация распознавания команд казахской речи на основе марковской модели. // Труды девятой международной азиатской школы-семинар «Проблемы оптимизации сложных систем» 2013.

18. Койбагаров К.Ч., Мусабаев Р.Р. Разработка лексико-морфологического парсера казахского языка // Труды 10-й международной Азиатской школы-семинар "Проблемы оптимизации сложных систем". Кыргызская Республика, Иссык-Кульская область, 2014 г.
19. Бекманова Г.Т. Некоторые подходы к проблемам автоматического словоизменения и морфологического анализа в казахском языке // Вестник Восточно-Казахстанского государственного технического университета им. Д. Серикбаева. – Усть-Каменогорск: 2009г. – № 4. – с. 192-197.
20. Жубанов А.Х. Основные принципы формализации содержания казахского языка. – Алматы, 2002. – 250 с.
21. Азарова И.В. Морфологическая разметка текстов на русском языке с использованием формальной грамматики AGFL. Кафедра математической лингвистики СПбГУ // <http://www.dialog-21.ru/Archive/2003/AzarovaAFGL.htm> (дата просмотра: 01.06.2016)
22. Барахнин В. Б., Федотов А. М. Построение модели фактографического поиска // Вестник НГУ. Серия: Информационные технологии. – 2013. – Том 11, Выпуск № 4. С. 16–27.
23. Chen P.P. The entity-relational model. Toward a unified view of data // ACM TODS. 1976. № 1. P. 9-36. / Рус. пер. Чен П. П.-Ш. Модель «сущность–связь» – шаг к единому представлению данных // СУБД. 1995. № 3. С.137-158.
24. Хайкин С. Нейронные сети. Полный курс, 2-е изд., испр.: Пер. с англ. – М.: ИОУ «И. Д. Вильямс», 2006.