НАЦИОНАЛЬНАЯ АКАДЕМИЯ НАУК КЫРГЫЗСКОЙ РЕСПУБЛИКИ

ИНСТИТУТ АВТОМАТИКИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

КЫРГЫЗСКО-РОССИЙСКИЙ СЛАВЯНСКИЙ УНИВЕРСИТЕТ

им. Б. Н. Ельцина

Диссертационный совет Д 05.18.579

На правах рукописи

УДК 519.688.004.912(575.2) (043.3)

КОЧКОНБАЕВА БУАЖАР ОСМОНАЛИЕВНА

РАЗРАБОТКА МОДЕЛЕЙ И АЛГОРИТМОВ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА ДЛЯ КЫРГЫЗСКОГО ЯЗЫКА

05.13.18- математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Бишкек - 2019

Диссертационная работа выполнена в институте природных ресурсов имени А. С. Джаманбаева Южного отделения Национальной академии наук Кыргызской Республики

Научный руководитель:	**Сатыбаев Абдуганы Джунусович** доктор физико-математических наук, профессор, заведующий кафедрой информационных технологий и управления Ошского технологического университета имени М. М. Адышева
Официальные оппоненты:	**Панков Павел Сергеевич** доктор физико-математических наук, член-корреспондент Национальной академии наук Кыргызской Республики, заведующий лабораторией вычислительной математики института математики Национальной академии наук Кыргызской Республики **Торобеков Бекжан** доктор технических наук, профессор, проректор по развитию и государственному языку Кыргызского государственного технического университета имени И. Раззакова
Ведущая организация:	**Ошский государственный университет, кафедра “Программирование”**, 723500, г. Ош, ул. Ленина, 331

Защита состоится 28 июня 2019 года в 14.00 часов на заседании диссертационного совета Д 05.18.579 при Институте автоматики и информационных технологий Национальной академии наук Кыргызской Республики и Кыргызско-Российском Славянском университете им. Б. Н. Ельцина, 720071, г. Бишкек, пр. Чуй, 265, ауд.346, сайт: www.iait.kg.

С диссертацией можно ознакомиться в библиотеках Национальной академии наук Кыргызской Республики по адресу: 720071, г. Бишкек, пр. Чуй 265 и Кыргызско-Российского Славянского университета им. Б. Н. Ельцина по адресу: 720000, г. Бишкек, ул. Киевская, 44 и на сайте Института автоматики и информационных технологий Национальной академии наук Кыргызской Республики по адресу: www.iait.kg. E-mail: gulsaat@mail.ru.

Автореферат разослан 27 мая 2019 г.

Ученый секретарь
диссертационного совета, к.ф-м.н _____Керимкулова Г. К.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы диссертации. На сегодняшнем этапе развития науки, техники и культуры предпочтение отдается процессам обработки информации, которые занимают лидирующие позиции в процессе общественного производства и проникают во все сферы деятельности человека. Методы и средства обработки информации на естественном языке приобретают все большее значение - от простейших систем подготовки документов до информационно-поисковых систем, систем машинного перевода и программ общения на естественном языке.

Программы, связанные с обработкой текстов на естественном языке очень много, и глубина их проникновения в структуру текста также различна.

В зависимости от необходимой глубины проникновения в структуру текста, эти программы обычно работают с блоками, которые соответствуют всей или часто первой части следующей последовательности обработки текста: буквы входного текста — слова — фразы — смысл — ... — новый смысл — структура фраз — значения слов — буквы выходного текста. Первым шагом обработки естественно-языкового текста является блок опознающий различные формы слов — блок морфологического анализа.

Во многих программах работающих обработкой естественного языка блок морфологического анализа является необходимой частью и это показывает актуальность задачи. Система требует, чтобы модуль морфологического анализа программы независимо от объема текста должна работать быстро и эффективно.

К таким создаваемым алгоритмам предявляются более жесткие требования, такие как потребность в меньшей памяти и высокой скорости, чем на известные алгоритмы. В этом направлении алгоритмы морфологического анализатора кыргызского языка широко не исследованы, поэтому создание формальной грамматики, машинного перевода, экспертных систем на основе морфологического, синтаксического и семантического анализаторов остается **актуальной задачей**.

Связь темы диссертации с приоритетными научными направлениями, крупными научными программами (проектами), основными научно-исследовательскими работами, проводимыми образовательными и научными учреждениями. Работа выполнена в институте природных ресурсов имени А. С. Джаманбаева Южного отделения Национальной академии наук Кыргызской Республики в рамках инициативной научно-исследовательской работы.

Цель исследования. Целью исследования диссертационной работы является создание математической модели автоматического морфологического анализа кыргызского языка, который оптимизирует обращение к оперативной

памяти и включающий лигвистические модели, а также разработка эффективно работающих на персональных компьютерах принципов, алгоритмов и программ на основе предложенной модели и использование этого модуля в прикладных программах, связанных с обработкой текста.

Задачи исследования:

- обзор существующих программ морфологического анализатора, которые эффективно работают на современных компьютерах;
- создание морфологической таблицы и словаря для кыгызского языка;
- разработка структуры базы данных при решении задач морфологического анализа;
- разработка математической модели морфологического анализатора для кыргызского языка;
- на основе математической модели создание алгоритма программы полного и точного морфологического анализа. А также на основе модели исследование задач нормализации слов;
- разработка программы морфологического анализатора в среде Embarcadero RAD Studio XE3.

Научная новизна полученных результатов:

1. Разработана модель морфологического строения кыргызского языка, основанный на разбиении словоформ на определенное количество частей с использованием правил. На основе предложенной модели разработан алгоритм нахождения нормальной формы слова.
2. Впервые разработана структура словаря, обеспечивающая получить основы слова за одно обращение к памяти компьютера. Это достигается за счет создания временного массива для хранения и обработки отсортированных слов из базы данных. А также создан алгоритм поиска слов в таком словаре.
3. Создана математическая модель морфологического анализатора и на основе модели алгоритм программы «NLP». Реализованная программа использует морфологический словарь кыргызского языка, который состоит из 15 тыс. лексем.

Практическая значимость полученных результатов. В результате исследования строения естественного языка была создана модуль системы, осуществляющая автоматический морфологический анализ на персональном компьютере. Это программное обеспечение позволяет повысить надежность поисковых систем и систем управления документами, а также использует языковую статистику, текстовые фрагменты для конкретных ситуаций, таких как синтаксический анализ, семантический анализ, машинный перевод, экспертные системы и т. д.

Результаты работы внедрены в работу электронной библиотеки для поисковых программ Ошского гуманитарно-педагогического института, в

учебный процесс Ошского технологического университета имени М. М. Адышева, а также проверены и получены положительные отзывы от экспертов национальной комиссии по государственному языку при президенте Кыргызской Республики.

Экономическая значимость полученных результатов. В результате диссертационной работы разработана программа морфологического анализатора кыргызского языка и в ходе работы были рассчитаны экономические показатели и эффективность разработанной системы. С помощью формул рассчитали экономическую эффективность программы, который составляет 35465,52 сомов.

Основные положения диссертации, выносимые на защиту:

1. Концептуальная схема работы морфологического анализатора.
2. Алгоритм оптимизации морфологического анализа.
3. Фрейм-модель формирования грамматических форм.
4. Математическая модель морфологии кыргызского языка.
5. Алгоритм функционирования морфологического анализатора.

Личный вклад соискателя. Построенные математические модели морфологического анализатора и алгоритмы программ, а также результаты, имеющие научную новизну и представленные в данной работе, получены лично автором. Научные советы по построению морфологического анализатора для кыргызского языка получены от профессора, д.фил.н. Т. Садыкова. Научному руководителю Сатыбаеву А. Дж. принадлежат общая постановка задач и обсуждение полученных результатов.

Апробации результатов диссертации. Основные научные результаты работы, докладывались автором на IV международной конференции «Turklang 2016» (Бишкек, 2016), V международной конференции «Turklang 2017» (Казань, 2017), VI международной конференции «Turklang 2018», посвященной компьютеризации языков тюркских народов (Ташкент, 2018), на конференции посвященной 75-летию М. Тагаева (Бишкек, 2018), на республиканской научно-практической конференции «Физико-технические проблемы в образовании и науки» (Ош, 2018).

Полнота отражения результатов диссертации в публикациях. Основные результаты диссертации отражены в 13 работах. Из них 5 опубликованы в Российских журналах, которые входит в РИНЦ, 6 статьи опубликованы в журналах, рекомендуемых ВАКом Кыргызской Республики. А также получено свидетельство Кыргызпатента на программное обеспечение “NLP Морфологический анализатор”.

Структура и объем диссертации. Данная диссертационная работа включает введение в работу, основное содержание в пяти главах, заключения, список литературы из 93 наименований и 4 приложения. Основное содержание диссертации изложена на 137 страницах и содержит 9 таблиц, 33 рисунка.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении разъясняется актуальность темы диссертационной работы, приведены цели, задачи и научная новизна, практическая значимость результатов исследований, а также основные положения, вносимые на защиту.

Первая глава «Обзор литературы» демонстрирует важность морфологического анализатора для работы с входными данными. Для достижения поставленных целей изучены применяемые методы и ранее исследованные морфологические программы и анализированы их преимущества и недостатки. Определены задачи исследования.

Изучения в области обработки естественного языка проводились кыргызскими учеными Т. Садыковым, Н. Исраиловой и П. Панковым. Первые работы в этой области сделаны Т. Садыковым.

Казахские ученные А. А. Шарипбаев, Д. Р. Рахимова, У. А. Тукеев, Ж.М. Жуманов посвятили свои работы для создания корпуса казахского языка и Д. Р. Рахимова в своей работе разрботала алгоритм машинного перевода с русского на казахский язык. Здесь она использовала таблицу сравнений морфологии двух языков.

Крупные практические прикладные системы созданы научным коллективом российских ученных в главе Г. Г. Белоногова, М. Г. Мальковского, И. А. Большакова, Ю. Д. Апресяна, О. С. Кулагиной, И. А. Батманова, Д. Варга, В. Н. Волкова, Е. Р. Добрушиной, Х. Ф. Исхаковой, Е. А.Казакова и В. А. Тузова. В. А. Тузов свои работы посвятил к созданию формальной модели русского языка основанный на правилах. Многие работы русских ученых направлены на исследования флективных языков.

Ученые K. Altintaş, İ. Çicekli исследовали турецкий язык, Д. Сулайманов, А. Гатиатулин татарский, А. С. Тантуг и другие тюркменские языки. Ученые J. Hankamer, L. Karttunen, M. Koskenniemi, H. Trost исследовали немецкий и английские языки.

Во второй главе «Материалы и методы исследования» были представлены материалы и методы, используемые в решении поставленных задач.

Объект исследования: объектом исследования является методы морфологической обработки текстов на естественном языке.

Предмет исследования: изучение строения словоформ кыргызского языка, создание программы автоматического морфологического анализатора для обработки естественного текста; визуализация морфологических данных, с хорошим доступом к словарю, хранящегося на жестком диске; методы и алгоритмы морфологического анализа и нормализации слов;

Методы исследования: для решения поставленной задачи применены методы морфологического анализа. Использованы элементы моделирования для построения математических моделей, описывающих лингвистические

закономерности, а также методы объектно-ориентированного программирования. В частности, мы рассматривали метод структурных схем морфологического анализа текстов естественного языка, на примере кыргызского языка, где словоформы состоят из корней (основное лексическое значение слова) и последовательности аффиксов, где каждый из которых выражает одно или несколько грамматических, или словообразовательных значений.

Термины «корень» и «основа» будем использовать как синонимы, для того чтобы объединить близких по смыслу словоформ. Аффиксы делятся на несколько типов в зависимости от их позиции относительно корня слова. Различают словообразовательные и словоизменительные аффиксы. В первом случае получаем модифицированное лексическое значение, а во втором случае аффиксы выражают только грамматические значения слов (например, падеж существительного или время глагола). В настоящей работе нашей основной задачей является выделение корня и множества аффиксов из входной словоформы. В естественном языке имеется следующие три случая представления слов: приставка+основа+аффикс, основа+аффикс, приставка+основа. Таким образом, получаем, что словоформа состоит из основы и определённого множества аффиксов, которое в некоторых случаях может оказаться пустым $\varnothing$.

Пусть задан алфавит $\Sigma = \{\sigma_1, \sigma_2,...\}$, где σ_i — буква.

Определение 1. Словом будем называть конечную последовательность $\sigma = \sigma_{i1} \sigma_{i2} ...\sigma_{in}$, где $\sigma_{ij} \in \Sigma$. Пустое слово будем обозначать как ε.

Определение 2. Множество $\Sigma*$ — множество всех допустимых слов над алфавитом Σ. Определим структуру словоформы, для этого введём несколько обозначений. Основу как тип обозначим через r, а множество слов такого типа обозначим через R. Также рассмотрим множество типов аффиксов A, отвечающих за формирование окончания. Пусть число различных типов будет m, тогда $A = \{a_1, a_2,...,a_m\}$. Для $i = 1,...,m$ каждому типу a_i поставим в соответствие множество A_i его представителей (слов). Тип префиксов обозначим через p, а множество его представителей через P. Важно отметить, что существуют слова, тип которых не определяется однозначно, т. е. $P \cap A_1 \cap A_2 \cap ... \cap A_m \neq \emptyset$. Через a_0 будем обозначать тип слов, которые не принадлежат множеству $P \cup A_1 \cup ... \cup A_m \cup R$. В общем случае такая последовательность выглядит следующим образом: $(a_{i1}, a_{i2}, ..., a_{is}), a_{ij} \in A$. Для описания всей некорневой части слова необходимо учитывать префиксную часть, поэтому в начало этой последовательности при наличии приставки нужно добавить p. Получили такую последовательность $(p, a_{i1}, a_{i2}, ..., a_{is}), a_{ij} \in A$. Учитывая особенности кыргызского языка, мы можем утверждать, что количество таких последовательностей конечно.

С третьей по пятой главы посвящены результатам собственных исследований и их обсуждениям.

В третьей главе «Уменьшение обращения к памяти компьютера при словарном морфологическом анализе» предлагаются разработанные методы организации словаря, позволяющие с помощью одного обращения к данным в памяти компьютера получать необходимую информацию. Предлагается алгоритм создания словаря и поиск данных в этом словаре.

Известны по крайней мере три способа создания морфологического анализатора: (1) анализатор, основанный на словарь, (2) анализатор, основанный на грамматику без словаря и (3) анализатор на базе грамматики и словаря.

Как известно, кыргызский язык входит в группу агглютинативных языков, где новые слова образуются с помощью словообразовательных аффиксов, а их грамматические формы - с помощью аффиксов словоизменительных. Здесь к основе слова в соответствии с правилами сингармонизма прибавляются окончания разной огласовки. Например: тоо+Ø=тоо, тоо+нын=тоонун, үй+нын= үйдүн и т.д.

В кыргызском языке текстовые формы существительных образуются с помощью довольно строгих правил агглютинации морфем, как это показано в фрейм-моделе следующего вида (рис. 1). При этом узлы модели указывают на разные состояния морфологии существительного, а линии соединяющие эти узлы – на конкретные категории имен существительных. N - это основа исследуемого существительного.

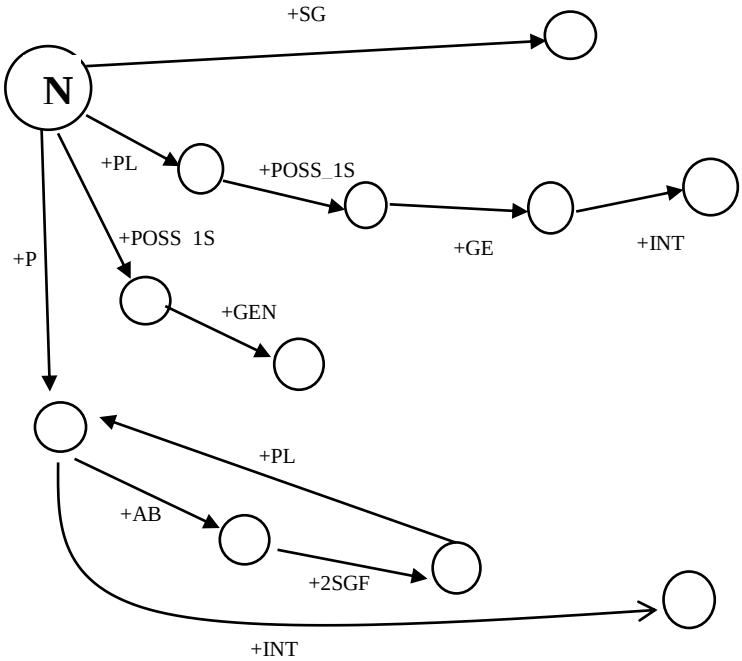

Рис. 1. Фрейм-модель формирования форм слов существительных

На базе этой фрейм-модели рассмотрим предикат функцию *word(x,y)*, где *x*-тип объекта, *y* – множество аффиксов. Здесь предикат *word* определяет отношение между основой и аффиксами. Рассмотренную модель можно представить следующим образом:

- *word(N,SG) - word(китеп, SG);*
- *word(N,P, POSS_1SG,GEN,INT) - word(китеп,P, POSS_1SG,GEN,INT);*
- *word(N, POSS_1SG,1GEN) - word(китеп, POSS_1SG,1GEN);*
- *word(N, PL, ABL, 2SGF, PL, INT) - word(китеп, PL, ABL, 2SGF, PL, INT).*

Таким образом, при рассмотрении принципов работы морфологического анализатора естественного текста необходимо учесть следующие этапы:

1. Разделение входного текста на грамматические формы слов.
2. Лемматизация текстовых форм слов в их словарную форму.
3. Разделение цепочки словоизменительных аффиксов на конкретные аффиксы.
4. Выявление морфологических признаков каждого из аффиксов.

Различают два подхода к решению задачи морфологического анализа словоформы, а именно: «справа налево» и «слева направо». При первом подходе делается попытка выделить конечную часть словоформы, похожую на множество аффиксов, и после проверяется наличие оставшейся начальной частив словаре. При втором подходе делается попытка найти в словаре некоторую начальную часть цепочки, а затем проверить, что оставшаяся правая часть образует возможный для данной основы комплекс аффиксов. При обоих подходах в случае неудачи поиска, приходится повторять другое разбиение словоформы.

Как правило, определение словарной информации является длительной операцией в морфологическом анализе. Точнее, появляется долговременная работа дискового пространства. Поэтому первым и вторым способом является необходимой задача по сокращению доступа к дисковой памяти как части полного цикла анализа слова.

Поскольку средняя длина слова в научно-техническом тексте составляет от 7 до 10 символов, подход «слева направо» может иметь 10 ссылок на словарь. Как правило, подход «справа налево» требует значительного количества ссылок на словарь, который является выбором этого конкретного метода при выполнении морфологического анализатора. Однако, по сравнению с направлением справа налево, подход «слева направо» имеет много преимуществ по сравнению с простотой реализации и возможностями.

Таким образом, задача максимального сокращения числа обращений к диску является актуальной. Конечно, в данном случае важным является создание подходящей базы данных и управления ими.

Построение базы данных

База данных — это множество данных, удовлетворящих потребностям пользователя. Эти данные сортируются и хранятся в виде таблицы, которым управляет система управления базами данных.

В данное время имеется множество систем управления базами данных:

SQL, MySQL, Oracle, Access. Работать с большим объемом данных всегда трудно и разные системы имеют свои преимущества. В нашем случае для обработки естественного языка и проверки предложенного алгоритма была создана в среде Embarcadero RAD Studio тестирующая программа. Здесь мы использовали среду Access для работы с данными.

У нас имеется три таблицы (основы, аффиксы и части речи) и связи между этими таблицами (рис.2):

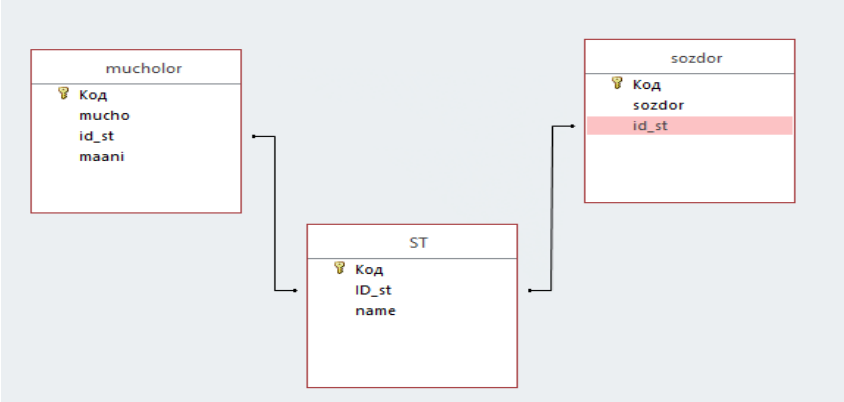

Рис. 2. Схема данных

Как видно из схемы данных, мы будем работать с тремя отдельными таблицами, которые связаны между собой ключевым полем it_st.

Оптимизация алгоритма морфологического анализа

Для оптимизации работы системы или сокращения числа обращений к диску мы использовали алгоритм следующего вида:

1. Выделяем из базы данных слова, которые начальные две буквы совпадают с начальными двумя буквами входного анализируемого слова по методу “слева направо”.
2. Создаем виртуальный массив для сохранения данных.
3. Находим основу слова используя метод “справа налево”.
4. Чтобы определить грамматическую категорию аффиксов будем использовать таблицу “mucholor”из базы данных. Количество итераций зависит от длины окончания.
5. В качестве результата получим текстовую форму фрейм-модели, который представлен на рисунке 1.

Как правило, морфемы как семантическая единица языка составляют основу словоформ и лексем. В кыргызском языке имеется четыре основные группы аффиксов. Для нахождения основы слова используются нижеследующие множества аффиксов.

Обозначим через P_i, *где* $i = 1, 2, 3, 4$, множества аффиксов.
Терминалы указывают на следующие конкретные множества:
P_1 – множества трехзначных аффиксов;
P_2 – множества аффиксов принадлежности;
P_3 – множества аффиксов указывающих личности;
P_4 – множества падежных аффиксов;
Q –предикат с одним аргументом; W – множество нормальных форм слова; N – существительное; V – глагол.
Если $x \in P_i$, то для всех $i = 1, \ldots, 4$ элементов получим $P_i(x)$.
Если слово $x \in W$, то обозначим через $W(x)$.
Если слово $x \in Q$, то обозначим $Q(x)$.
Тогда наши правила *A–H* изменяются по следующим формулам.
Допустим любое слово z составлен из букв $x_0 + x_1 + x_2 + \ldots + x_k$. ($x_i$ – максимальное количество букв) $i = k$, $x = x_i$, то получим следующий алгоритм нахождения нормальной формы слова:

1-шаг

$$A = \begin{cases} P_4(x) \to Q(z\backslash x), & z = z\backslash x \in N, x_{i-n} + \cdots + x_{i-1} + x_i \neq P_1 \\ & n = length(P_1) \\ P_1(x) \to Q(z\backslash x), & z = z\backslash x \in N, \\ & x_{i-2} + x_{i-1} + x_i \neq P_4 \end{cases}$$

2-шаг

$$B = \begin{cases} P_2(x) \to Q(z\backslash x), z = z\backslash x \in N, x_{i-2} + x_{i-1} + x_i \neq P_1 \\ \qquad x_{i-2} + x_{i-1} + x_i \neq P_4 \\ P_1(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \qquad x_{i-2} + x_{i-1} + x_i \neq P_4 \end{cases}$$

3-шаг

$$C = \begin{cases} P_4(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \qquad x_{i-2} + x_{i-1} + x_i \neq P_1 \\ P_3(x) \to Q(z\backslash x), z = z\backslash x \in V \\ P_1(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \qquad x_{i-n} + \cdots + x_{i-1} + x_i \neq P_3 \\ \qquad n = length(P_3) \end{cases}$$

4-шаг

$$D = \begin{cases} P_2(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i \neq P_1 \\ P_3(x) \to Q(z\backslash x), z = z\backslash x \in V, \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i = P_4 \\ P_1(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i \neq P_3 \end{cases}$$

5-шаг

$$E = \begin{cases} P_1(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \qquad x_{i-2} + x_{i-1} + x_i \neq P_3 \\ P_2(x) \to Q(z\backslash x), z = z\backslash x \in V, x_{i-n} + \cdots + x_{i-1} + x_i = P_1 | \\ \qquad x_{i-n} + \cdots + x_{i-1} + x_i \neq P_4 \end{cases}$$

6-шаг

$$F = \begin{cases} P_4(x) \to Q(z\backslash x), \qquad z = z\backslash x \in N, \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i \neq P_1 \\ P_3(x) \to Q(z\backslash x), z = z\backslash x \in V, \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i = P_4 \end{cases}$$

7-шаг

$$G = \begin{cases} P_2(x) \to Q(z\backslash x), z = z\backslash x \in V, \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i = P_1 \\ P_3(x) \to Q(z\backslash x), z = z\backslash x \in V \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i = P_4 \end{cases}$$

8-шаг

$$H = \begin{cases} P_3(x) \to Q(z\backslash x), z = z\backslash x \in V \\ \quad x_{i-n} + \cdots + x_{i-1} + x_i = P_4 \\ P_1(x) \to Q(z\backslash x), z = z\backslash x \in N, \\ \quad x_{i-2} + x_{i-1} + x_i \neq P_3 \end{cases}$$

После этого шага останавливается проверка аффиксов, если проверка неуспешная, то возвращаемся к первому шагу. Таким образом, получим основу словоформы.

В четвертой главе «Математическая модель морфологического анализа кыргызского языка» описывается формальная математическая модель кыргызского языка с особенностями морфологических строений, а также структура программы, разработанный автором, на основе данной модели (рис.3). Приведены алгоритмы морфологического анализатора и нормализации слов.

Морфологические категории. Их перечень определяется следующими категориями:

1. Имя существительное – Noun:

Категория числа – Number

1. Единственное число – singular.
2. Множественное число – plural.

Тэги:

1. **SG <=>∅**
2. **PL <=> ЛАр**

Категория притяжательности – Possessive

Единственное число – singular:

1) первое лицо единственного числа - 1st person singular possessive (‘my’);
2) второе лицо единственного числа - 2nd personsingularpossessive (‘your’);
3) второе лицо единственного числаласк. - 2nd person sing.po**ss.** formal (‘your’);
4) третье лицо единственного числа - 3nd person singular possessive (‘his/her/its’);
Множественный падеж – plural:
5) первое лицо множественного числа - 1st person plural possessive (‘our’);
6) второе лицо множественного числа - 2nd person plural possessive (‘your’);
7) второе лицо множественного числа ласк. - 2nd person pl.po**ss.** formal (‘your’);
8) третье лицо множественного числа - 3nd person plural possessive (‘their’) и т.д.

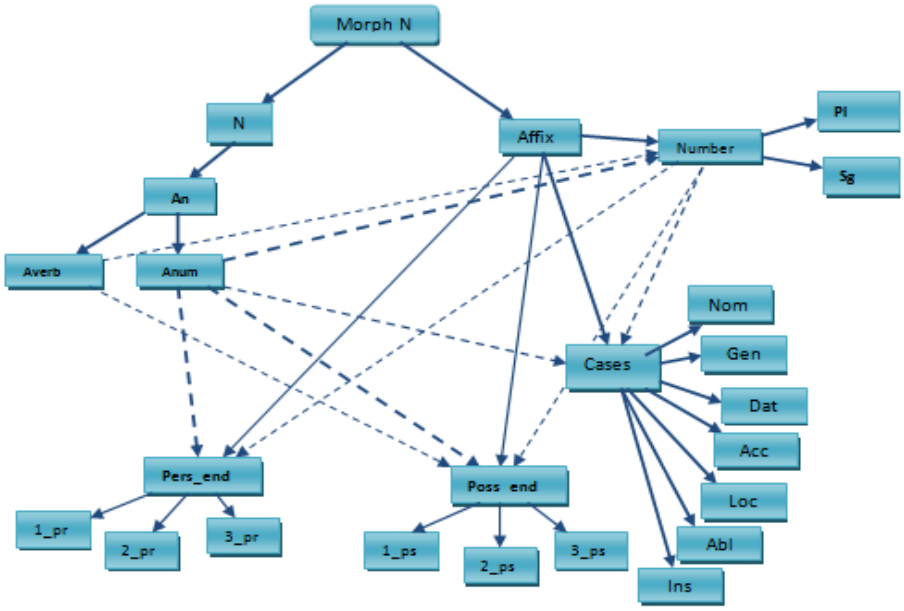

Рис. 3. Модель создания существительных

Математическое моделирование морфологического анализа кыргызского языка

Мы обозначаем словоформу в любом агглютинативном языке строкой $S_n = x_1 x_2 \dots x_n$, где x_i *(i=1,2,…,n)* является членом соответствующего алфавита А, а *n* является количеством букв (то есть длиной строки). В нашем исследовании мы используем кыргызский алфавит, который состоит из 36 букв и знака подчеркивания _ для пустого символа следующим образом:

А={а,б,в,г,д,е,ё,ж,з,и,й,к,л,м,н,ң,о,ө,п,р,с,т,у,ү,ф,х,ц,ч,ш,щ,ъ,ы,ь,э,ю,я,э, ’_’}

и мы ввели следующие обозначения S_n для обозначения подстрок любой строки $1 \le i \le j \le n$:

$$S_n[i{:}j] = x_i\, x_{i+1} \dots x_j$$

$$S_n[{:}j] = x_1\, x_2 \dots x_j$$

$$S_n[i{:}] = x_i\, x_{i+1} \dots x_n$$

Исходя из наших обозначений, специальная подстрока $S_n[i{:}\,i+1] = x_i\, x_{i+1}$ обозначается упорядоченной парой букв $(x_1,x_2)_i$, где субиндекс

$i(i=1,2,\ldots,n-1)$, указывает начальную позицию упорядоченной пары в этой строке $x_1=x_i$, $x_2=x_{i+1} \in A$.

Для $i=n$ упорядоченная пара формируется добавленным пробелом как $(x_n, '_')_{i=n}$. Таким образом, любая строка $S_n = x_1x_2 \ldots x_n$ имеет n упорядоченную пару в нашем исследовании.

Для заданной упорядоченной пары букв $(x_1,x_2)_j$ которая может появляться в позиции $1 \leq j \leq n_{max}$в любой форме кыргызского слова (где n_{max} максимальная длина слова в кыргызском языке) и данная конкретная форма слова обозначается как $S_n = x_1x_2 \ldots x_n$, где $n \geq j$, обозначение $(x_1,x_2)_j \in S_n$ указывает, что существует упорядоченная пара $(x_1,x_2)_i$ в позиции i $(1 \leq i \leq n)$ в S_n при условии, что $(x_1, x_2)_i = (x_1,\ x_2)_j$ *для* $i=j$. Наконец, мы определяем еще два символа, а именно $g_m=S_n[:m]$ *и* $e_m=S_n[m:]$ чтобы представить любую словесную форму в виде упорядоченной пары из двух подстрок $S_n^m = (g_m, e_m)$ для всех $1 \leq m \leq n$.

Предположим, что множество L будет набором всех возможных упорядоченных пар букв $(x_1,x_2)_i$ которое может появляться в любой кыргызской словесной форме для позиций $i=1,\ldots n_{max}$. Тогда L будет пробным пространством и может быть определено следующим образом:

$$L = \{(x_1,\ x_2)_i | x_1,\ x_2 \in A\ and\ 1 \leq i \leq n_{max}\}$$

И далее предположим, что множества G_k , E_k и T_k, *где* G_k , E_k , $T_k \subset L, 1 \leq k \leq n_{max}$ представляют множества, определенные следующим образом:

$$G_k = \{(x_1,\ x_2)_i | i = k\ and\ (x_1,\ x_2)_i \in g_m\ and\ 1 \leq m \leq n_{max}\}$$

$$E_k = \{(x_1,\ x_2)_i | i = k\ and\ (x_1,\ x_2)_i \in e_m\ and\ 1 \leq m \leq n_{max}\}$$

$$T_k = \{(x_1,\ x_2)_i | i = k,\ h_1 = s_n[k{:}\,k],\ h_2 = s_n[k+1, k+1],\ 1 \leq i \leq n_{max}\}$$

Таким образом, для каждой упорядоченной пары $(x_1, x_2)_i$ в позициях $i=1,2,\ldots,n$ любой заданной словоформы, обозначенной через $S_n = x_1x_2 \ldots x_n$ можно определить вероятности нахождения в вышеуказанных три множества следующим образом:

$$Pr(s_n[i{:}\,i+1] \in G_i) = Pr((x_1,\ x_2)_i \in G_i) = P_G((x_1,\ x_2)_i) \quad (1)$$

$$Pr(s_n[i{:}\,i+1] \in E_i) = Pr((x_1,\ x_2)_i \in E_i) = P_E((x_1,\ x_2)_i) \quad (2)$$

$$Pr(s_n[i{:}\,i+1] \in T_i) = Pr((x_1,\ x_2)_i \in T_i) = P_T((x_1,\ x_2)_i) \quad (3)$$

где, уравнение (1) показывает, что упорядоченная пара $(x_1, x_2)_i$ находится в основной части заданной формы слова, аналогично уравнение (2) показывает, что упорядоченная пара $(x_1, x_2)_i$ находится в аффиксной части данной формы слова и, наконец, уравнение (3) показывает, что упорядоченная пара $(x_1, x_2)_i$ находится между частью основы и аффиксной частью данной формы слова (то есть, x_1 - последняя буква части стебля, а x_2 - первая буква части аффикса).

Ввиду того, что слова кыргызского языка состоят из корня и аффиксов, слово обозначим как S, тогда в качестве функции их определим так:

$$S = R + \sum_{i=0}^{m} U_i\,, (m \leq 8) \qquad (4),$$

здесь, S - линейная функция, R - основа слова, Um - словоизменительные аффиксы.

Из формулы (4) видно, что S зависит от основы, словообразовательных и словоизменительных аффиксов.

Словоизменительные аффиксы могут достичь до восьми, иначе говоря

$$\sum_{i=0}^{8} U_i = U_0 + U_1 + U_2 + \cdots + U_8, \ (5)$$

Определение 1: Если *Km=∅, Um=∅* , то S функция будет равна корню слова, и вводимое слово не разделится на морфемы.

Если длина искомой словоформы находим через *l=length(S),* то множество найденных словоформ из словаря обозначим через m и получим сегментационную матрицу l × m.

$$\begin{array}{c} l \\ m\left\{\begin{array}{c} \underbrace{1\ \ 1\ \ a_{31}a_{41}\ \ \dots\ \ a_{l1}}_{i\geq l} \\ \\ 1\ \ 1\ \ a_{32}a_{42}\ \ \dots\ \ a_{l2} \\ \dots \\ 1\ \ 1\ \ a_{3m}a_{4m}\ \ \dots\ \ a_{lm} \end{array}\right\} \end{array}$$

В данной матрице каждая строка одна словоформа и искомая основа слова находим сравнивая длины строк. Из матрицы получим следующую систему неравенств:

$$\begin{cases} x_1 + x_2 + x_3 + \cdots + x_l \leq l \\ \dots \\ x_1 + x_2 + x_3 + \cdots + x_{l-i} \leq l - i \end{cases}$$

Решая систему неравенств получим множество элементов одной строки, где содержится наибольшее количество единиц.

Например, при анализе слова “*абалы”* из словаря основ находим 6 элементов массива. После сегментации элементов массива на компьютере получим следующую матрицу:

$$M=\begin{vmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{vmatrix}$$

Анализируя элементы матрицы получим элементы 4 и 5 строк для дальнейшего анализа.

$$M=\begin{vmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{vmatrix}$$

Из элементов этих двух строк создаем одномерный массив и получим одномерный массив из длин этих строк $L=\{l_1, l_2\}$. Сравнивая длины и единицы получим искомую основу. В даном случае количество слов анализа равно двум. Иногда этот показатель достигает на много больше.

Приближенное значение полученного массива находим через итерацию и это имеет следующий код:

```
For i:=1 to length(L) do
If min>l[i] then min:=l[i];
```

В результате этого механизма получим строку минимальную по длине. В нашем случае этот элемент M[5] т.е. выполняется следующий фрагмент кода:

```
ss2:=trim(Edit1.Text);
d:=strtoint(edit3.Text);
edit2.Text:='';
1:l:=length(ss2);
for i:=1 to 46 do
if (ss2=m[i]) or (ss2=(copy(m[i],1,length(m[i])-1))+'б')
or (ss2=(copy(m[i],1,length(m[i])-1))+'г') then
begin
edit2.Text:=m[i];
form1.RichEdit3.Lines.Add(mucho.edit2.Text);
case strtoint(m3[i]) of
1:form1.RichEdit3.Lines.Add('зат атооч');
2:form1.RichEdit3.Lines.Add('сын атооч');
3:form1.RichEdit3.Lines.Add('сан атооч');
4:form1.RichEdit3.Lines.Add('ат атооч');
5:form1.RichEdit3.Lines.Add('этиш');
6:form1.RichEdit3.Lines.Add('тактооч');
7:form1.RichEdit3.Lines.Add('сырдык сөз');
8:form1.RichEdit3.Lines.Add('тууранды сөз');
9:form1.RichEdit3.Lines.Add('жандооч');
10:form1.RichEdit3.Lines.Add('байламта');
11:form1.RichEdit3.Lines.Add('кызматчы сөз');
end;end;
ss3:=ss3+copy(ss2,l,l);
ss2:=copy(ss2,1,l-1);
if edit2.Text='' then goto 1
else if (m[i]='') then
begin
```

timer1.Enabled:=true; end; aff:='';
for i:=1 to length(ss3)-1 do
aff:=aff+ss3[length(ss3)-i];edit4.Text:=aff;

Алгоритм морфологического анализатора

Соответствующие блок схемы алгоритма представлены на рис. 4 и 5.

Можно рассмотреть 3 ступени проведения морфологического анализа:

1. Определение только грамматического значения слова.
2. Определение основы слова.
3. Определение грамматических значений и основы слова.

Развернутое или неполное исследование морфологического разбора зависит от поставленной задачи.

Морфологический анализ является начальной ступенью различных задач, связанных с естественным языком, и поэтому его точное выполнение имеет большое значение.

Методы морфологического анализа можно разделить на 3 типа:

- анализировать со словарем аффиксов;
- анализировать с помощью словаря аффиксов и основ;
- анализировать с помощью словаря системы слов.

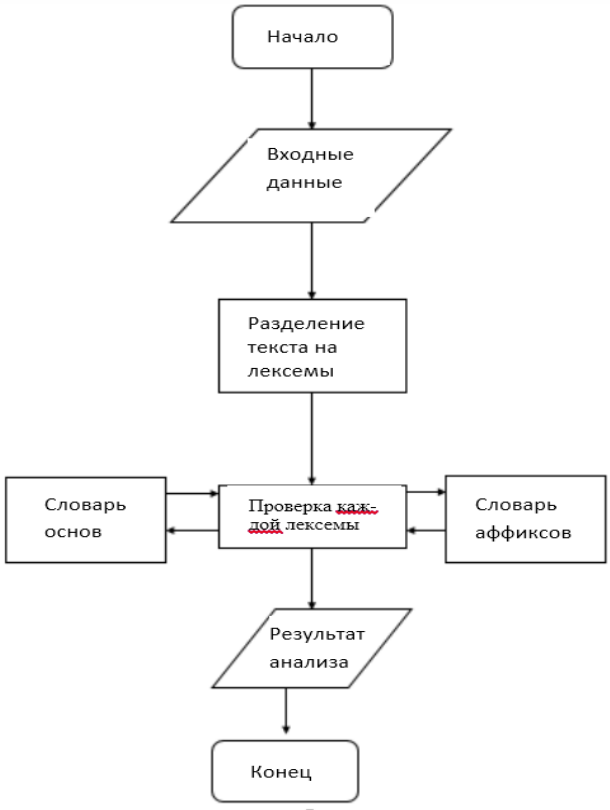

Рис. 4. Алгоритм морфологического анализа естественного текста

Рис. 5. Алгоритм модуля разделения текста на лексемы

В пятой главе «Программное обеспечение морфологического анализатора» получаем результаты тестирования на работоспособность разработанной программы.

Морфологический анализатор берет слово в качестве входных данных и вырабатывает корень, а его грамматические функции - как результат. На основе морфологического анализа в среде RAD Studio XE3 была составлена система анализатора NLP. Выбрано объектно-ориентированная среда RAD Studio XE3. Среда поддерживает тип unistring, который поддерживает алфавит кыргызского языка.

Система состоит из базы данных и интерфейса пользователя, модуля морфологического анализатора и статистического анализа.

Составленная в среде RAD Studio XE3 программное обеспечение состоит из процедур и функций, которые составляют 800 строк и занимает 15,7 Мб компьютерной памяти. При работе с базами данных используется 16 Мб памяти и для лингвистических таблиц расходуется 40 Кб памяти.

Таким образом, прикладная программа по морфологическому анализу текстов естественного языка “NLP” состоит из 69 функций и 22 постоянных параметров. В начале программы была построена концептуальная схема системы (рис. 6).

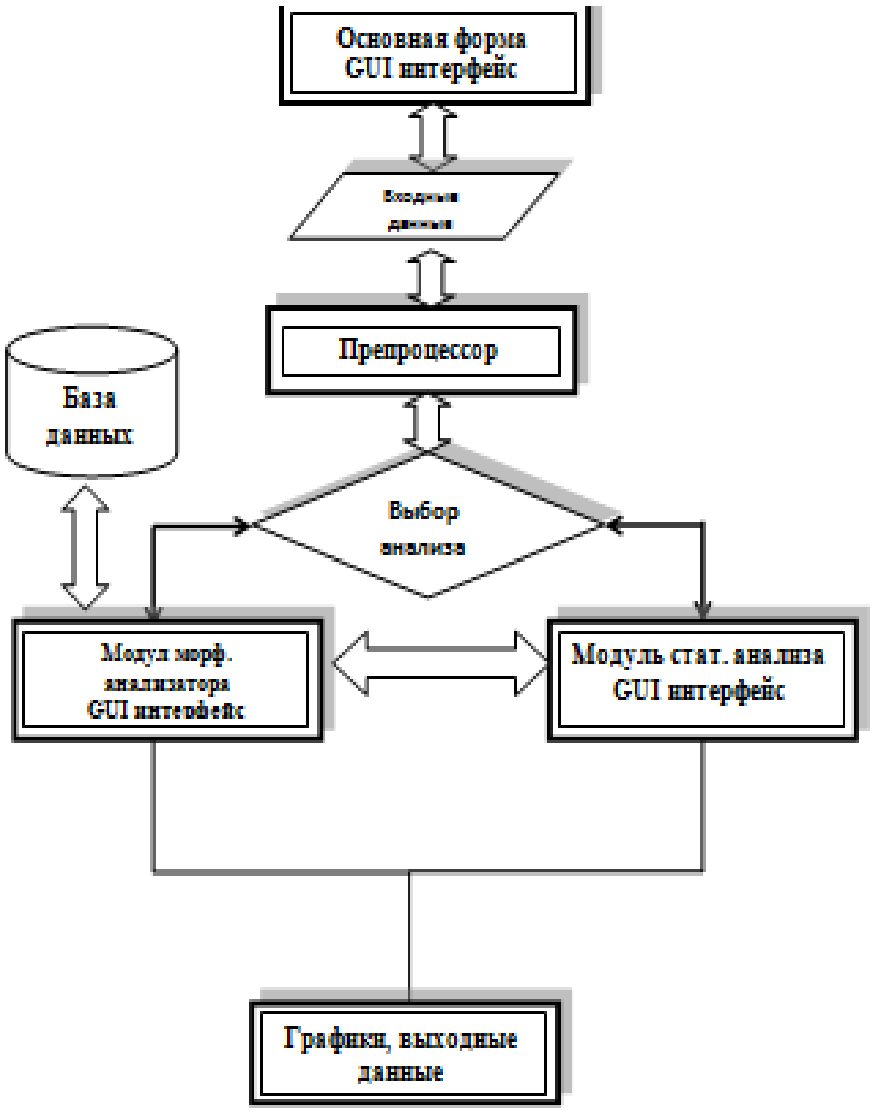

Рис.6. Концептуальная схема системы NLP

Тестирование системы

Специфика первой версии морфологического анализатора в том, чтобы максимум информации заложить в базе данных с относительно простой программной частью. Программная часть реализована в виде GUI интерфейса (рис. 7), который позволяет производить запросы к базе данных.

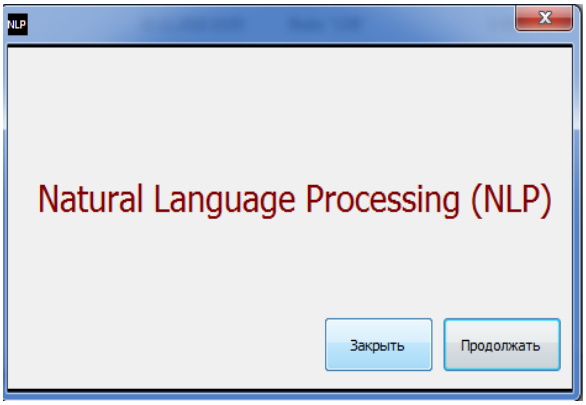

Рис. 7. Интерфейс системы

Таким образом, весь процесс морфологического анализа представляет собой поиск в базе данных элементов, удовлетворяющих заданным параметрам. База данных морфологического анализатора состоит из двух словарей: словарь основ и словарь аффиксов. В словаре основ все основы классифицированы по морфологическим типам, а в словаре аффиксов хранятся множества типов окончаний. Окончания представлены в виде множества аффиксов, образованных по морфонологическому правилу кыргызского языка (рис. 8).

Код	mucho	
1	лар	көптүк, PL
2	лер	көптүк, PL
3	лор	көптүк, PL
4	лөр	көптүк, PL
5	дар	көптүк, PL
6	дер	көптүк, PL
7	дер	көптүк, PL
8	дор	көптүк, PL
9	дөр	көптүк, PL
10	тар	көптүк, PL
11	тер	көптүк, PL

а)

код	sozdor	id_st
1	аалам	1
2	ааламдык	2
3	аалаш	5
4	аалим	1
5	аалы	1
6	алым	1
7	аамыят	1
8	аарчы	5
9	аарчыл	5
10	аарчын	5
11	аары	1
12	аба	1
13	абаз	1
14	абай	1

б)

Рис. 8. а) база данных аффиксов; б) база данных основ

С другой стороны, эти цепочки образованы словоизменительными аффиксами, которые в агглютинативных языках иногда имеют бесконечную длину. В нашем случае при создании базы данных приняли ограничение по заполнению цепочек окончаний, состоящих не более чем из восьми аффиксов, что является обоснованным со статистической точки зрения.

Механизм работы программы морфологического анализа заключается в следующем. Программа морфологического анализа проверяет возможность

получения аффиксальных цепочек на основе правил следования алломорфов, а также соответствие типа, получаемой основы необходимым для используемых алломорфов морфонологическим признакам. Вся требуемая для работы программы информация находится в оперативной памяти, которая загружается при запуске программы. Таким образом, отсутствует обращение к сетевой базе данных, что способствует увеличению скорости обработки анализируемых данных.

Разработан модуль анализа кыргызского языка, который выполняет разбиение текста на кыргызском языке на слова, и для каждого слова устанавливает нормальную форму и морфологические признаки. Результаты анализа текстов на кыргызском языке отображаются в графическом интерфейсе (рис. 9).

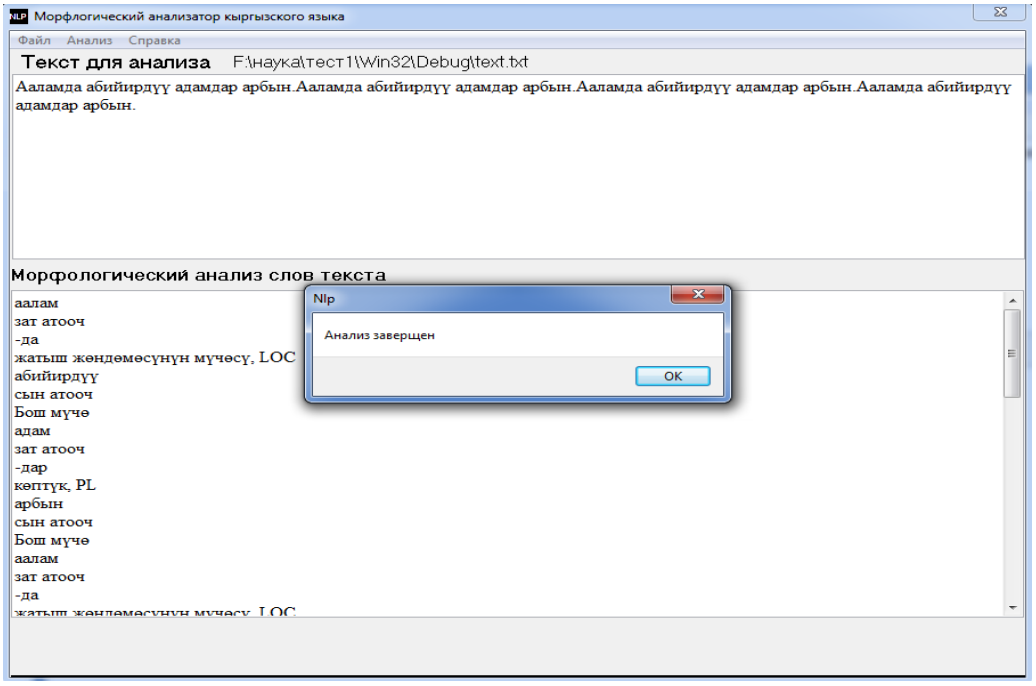

Рис. 9. Результат работы программы анализатора

А также система имеет модуль статистического анализа, где вычисляется частота выполнения проверки и разбиения каждого слова (рис. 10).

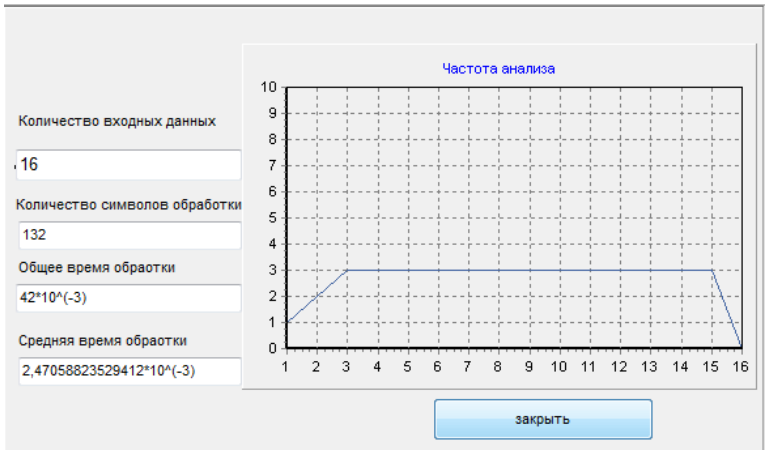

Рис. 10. Результат статистического анализа

Создание морфологического анализатора показали, что кыргызский язык как и все тюркские языки является агглютинативным языком и лексемы языка состоят из основы и множества аффиксов.

Результаты работы программы NLP показали положительные ответы и модули системы могут быть использованы для информационно-поисковых систем и машинного перевода как первые этапы обработки текстов естественного языка.

В заключении приведены полученные в диссертационной работе основные научно-практические результаты.

В приложении содержится листинг разработанного программного кода, акты внедрения и свидетельство Кыргызпатента на программное обеспечение.

ВЫВОДЫ

После научного исследования получили следующие выводы:

1. Созданы модели отображающие морфологические строения агглютинативного естественного языка, разработанная на персональных компьютерах нового поколения. Программа является эффективной, из-за быстродействия и небольшой потребности в оперативной памяти компьютера.
2. Разработана математическая модель морфологического анализатора для кыргызского языка.
3. С использованием предложенной модели разработан алгоритм полного морфологического анализа и нормализации слов, а также программный продукт. Предложенные алгоритмы и модули составляют универсальную морфологическую систему с изменением базы данных.
4. Созданы системы морфологических данных для кыргызского языка, а также словарь кыргызского языка, включающий около 15000 лексем.
5. Из-за потребностей программы морфологического анализатора предложена структура базы данных. При управлении с данными программа должна обращаться в память компьютера один раз и создать массив необходимых слов для обработки полученных данных, обеспечивая этим эффективность использования памяти компьютера.
6. Создано программное средство “NLP”, реализующее автоматический морфологический анализ над входными данными. Программа составлена в среде программирования Embarcadero RAD Studio XE3.

Практические рекомендации

Результаты полученных работ применяются как начальный модуль при создании прикладных программ по обработке текстов естественного языка.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. **Кочконбаева, Б. О.** Automatic processing of text in natural language [Текст] / Б. О. Кочконбаева, А. Алдосова // Бюллетень науки и практики. – 2018. – Т. 4, № 7. – С. 216-221.

2. **Кочконбаева, Б. О.** Алгоритм синтаксического анализатора для машинного перевода текстов [Текст] / Б. О. Кочконбаева // Труды V Междунар. науч.-практ. конф. информатизация общества. – Астана, 2016. – С.92-95.

3. **Кочконбаева, Б. О.** Защита информации с помощью криптографических методов [Текст] / Б. О. Кочконбаева, Н. Р. Абдыраева // Изв. ОшТУ. – 2010. – № 2. – С.183-186.

4. **Кочконбаева, Б. О.** Лексический анализатор естественного текста [Текст]/ Б. О. Кочконбаева, Н. Р. Абдыраева // Изв. ОшТУ. – 2014. – С.207-209.

5. **Кочконбаева, Б. О.** О морфологическом анализе в приложениях автоматической обработки текста (АОТ) [Текст] / Б. О. Кочконбаева // Бюллетень науки и практики. – 2018. – Т. 1, № 12.– С.608-612.

6. **Кочконбаева, Б. О.** Об оптимизации алгоритма морфологического анализа [Текст] / Б. О. Кочконбаева, Т. Садыков. – Ташкент, 2018. – С. 293-299

7. **Кочконбаева, Б. О.** Компьютерная обработка естественного языка [Текст] / Б. О. Кочконбаева, Н. Р. Абдыраева // Изв. ОшТУ. – 2015. – С.86-89.

8. **Кочконбаева, Б. О.** Табигый тилдеги тексттерди орус тилинен кыргыз тилине машиналык которууда создорду анализдоонун алгоритмин тузуу [Текст] / Б. О. Кочконбаева // Изв. КТУ им. Раззакова. – 2016. – № 2(38). – С.52-54.

9. **Кочконбаева, Б. О.** Кыргыз тили учун создун негизин аныктоо модели [Текст] / Б. О. Кочконбаева // Изв. ОшТУ. – 2018. – № 1. – С.24-30.

10. **Кочконбаева, Б. О.** Улуттук корпус учун морфологиялык белгилоолор [Текст] / Б. О. Кочконбаева, Т. С. Садыков, Б. Ш. Шаршембаев // Вестн. КРСУ. – 2018. – Т. 18, № 1. – С.91-95.

11. **Кочконбаева, Б. О.** Модель морфологического анализа кыргызского языка [Текст] / Б. О. Кочконбаева, Т. С. Садыков // Издательство Академии наук Республики Татарстан – Казань, 2017. – С.135-155.

12. **Кочконбаева, Б. О.** Математическое моделирование и алгоритм морфологического анализа кыргызского языка [Текст] / Б. О. Кочконбаева, А. Дж. Сатыбаев // Бюллетень науки и практики. – 2019. – Т. 5, № 3. – С. 220-224.

13. **Кочконбаева, Б. О.** Тестирование программы морфологического анализатора естественного языка [Текст] / Б. О. Кочконбаева, А. Дж. Сатыбаев //Бюллетень науки и практики. – 2019. – Т. 5, № 3. – С. 215-219.

14. **Кочконбаева, Б. О.** Программа для ЭВМ «Natural Language Processing. Морфологический анализатор кыргызского языка» [Текст] / Свидетельство КР, №537-Кыргызпатент, 19.12.2018.

Көчкөнбаева Бүажар Осмоналиевнанын 05.13.18 - математикалык моделдөө, сандык ыкмалар жана программалар комплекси адистиги боюнча «Кыргыз тили үчүн морфологиялык анализатордун моделдерин жана агоритмдерин иштеп чыгуу» аттуу темасында аткарылган диссертациясынын

РЕЗЮМЕСИ

Ачкыч сөздөр: морфологиялык анализатор, машиналык которуу, стемминг, морфологиялык талдоо, лемматизация, аффикстер, сөз формасы, сөздүн нормалдык формасы.

Изилдөө объектиси: табигый тилдеги текстерди морфологиялык иштеп чыгуунун ыкмалары изилдөөнүн объектиси болуп эсептелет.

Изилдөөнүн предмети: текстти иштеп чыгуунун автоматтык морфологиялык анализинин программасын түзүүдө, аны формалдык түрдө кароо үчүн керек болгон кыргыз тилинин морфологиялык түзүлүшүн изилдөө; сөздүктү сунуштоонун ыкмалары жана компьютердин эсинде сакталган сөздүккө жеткиликтүүлүктү ылдамдатууга байланыштуу морфологиялык маалыматтар;
Морфологиялык анализдин жана сөдөрдү негизин табуунун ыкмалары жана алгоримдери.

Иштин максаты: морфологиялык анализатордун алгоритмдерин жана моделдерин түзүү.

Изилдөө ыкмалары: алдыда коюлган маселени чечүүдө морфологиялык анализдин ыкмалары, лингвистикалык мыйзам ченемдүүлүктөрдү туюндуруучу математикалык моделдөө элементтери, ошондой объектке багытталган программалоонун ыкмалары колдонулган.

Аппаратура: ноутбук Intel Core i3, Embarcadero RAD Studio XE3

Иштин негизги натыйжалары: морфологиялык анализдин математикалык моделдери жана алгоритмдери ишелип чыкты, ошондой эле морфологиялык анализатордун автоматташтырылган системасы жана көп колдонулуучу сөздөрдүн сөздүгү түзүлгөн.

Изилдөөнүн натыйжаларын колдонуу: иштелип чыккан морфологиялык анализатордун системасы М.М. Адышев атындагы Ош технологиялык университетинин окуу процессинде колдонууга киргизилди жана мамлекеттик Ош педагогикалык институтунун электрондук библиотекасына маалымат издөөчү модул катары кириштелди. Ошондой эле Кыргыз Республикасынын Президентине караштуу Мамлекеттик тил боюнча улуттук комиссиясынын эксперттери тарабынан жактырылды.

Колдонуу тармагы: изилдөөнүн натыйжалары жана иштелип чыккан система машиналык которуу, эксперттик системаларда, окутуу жана үйрөтүүчү системаларда базалык модуль катары колдонулат.

РЕЗЮМЕ

диссертации Кочконбаевой Буажар Осмоналиевны на тему: "Разработка моделей и алгоритмов морфологического анализатора для кыргызского языка" на соискание ученой степени кандидата технических наук по специальности 05.13.18 - математичексое моделирование, численные методы и комплексы программ

Ключевые слова: морфологический анализатор, машинный перевод, стемминг, морфологический анализ, лемматизация, аффиксы, словоформа, нормальная форма слов.

Объект исследования: методы обработки текстов естественного языка.

Предмет исследования: изучение строения словоформ кыргызского языка, создание программы автоматического морфологического анализатора для обработки естественного текста; визуализация морфологических данных, с хорошим доступом к словарю, хранящегося на жестком диске; методы и алгоритмы морфологического анализа и нормализации слов;

Цель исследования: разработка автоматизированного морфологического анализатора.

Методы исследования: при решении поставленных задач в работе использованы методы морфологического анализа. Применены элементы моделирования для построения математических моделей, описывающих лингвистические закономерности, а также методы объектно-ориентированного программирования.

Аппаратура: ноутбук Intel Core i3, Embarcadero RAD Studio XE3

Полученные результаты и их новизна: разработаны математические модели и алгоритмы морфологического анализа, а также автоматизированная система морфологического анализатора и словарь с часто используемыми словами.

Использование результатов исследования: автоматизированная система морфологического анализа внедрена в учебный процесс Ошского технологичексого университета имени М. М. Адышева и в электронную библиотеку Ошского государственного педагогического института в качестве модуля поиска информации. А также программа получила положительные отзывы от экспертов национальной комиссии по Государственному языку при Президенте Кыргызской Республики.

Область применения: результаты исследования и разработанная система могут быть использованы в системах машинного перевода, экспертных системах, обучающих системах как базисный модуль.

SUMMARY

of the dissertation of Kochkonbaeva Buazhar Osmonalievna on the theme: "Development of models and algorithms of morphological analyzer for the Kyrgyz language" for the degree of candidate of technical sciences, specialty 05.13.18 - mathematical modeling, numerical methods and program complexes.

Keywords: morphological analyzer, machine translation, stemming, morphological analysis, lemmatization, affixes, word form, normal form of words.

Object of the research: natural language text processing methods.

Subject of research: study of the structure of word forms of the Kyrgyz language, the creation of an automatic morphological analyzer program for processing natural text; visualization of morphological data, with good access to the dictionary stored on the hard disk; methods and algorithms for morphological analysis and normalization of words;

Purpose of the research: development of an automated morphological analyzer.

Research methods: in solving the tasks in the work used analytical methods. The elements of modeling are used to build mathematical models that describe linguistic patterns, as well as methods of object-oriented programming.

Hardware: laptop Intel Core i3, Embarcadero RAD Studio XE3.

The results obtained and their novelty: developed mathematical models and algorithms for morphological analysis, as well as an automated morphological analyzer system and a dictionary with frequently used words.

Using the results of the study: an automated system of morphological analysis was introduced into the educational process of the Osh technological University after named M. M. Adyshev and the electronic library of the Osh State Pedagogical Institute as a module for information retrieval. As well as the program received positive feedback from experts of the national commission on the State language under the President of the Kyrgyz Republic.

Scope: the research results and the developed system can be used in machine translation systems, expert systems, training systems, as a basic module.

Кочконбаева Буажар Осмоналиевна

РАЗРАБОТКА МОДЕЛЕЙ И АЛГОРИТМОВ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА ДЛЯ КЫРГЫЗСКОГО ЯЗЫКА

Автореферат диссертации

Подписано в печать 17.05.2019 г.
Формат 60x84/16. Объем 1,25 п.л.
Бумага офсетная. Печать офсетная.
Тираж 40 экз.

720020, Бишкек, ул. Малдыбаева, 34 б,
Кыргызский Государственный университет строительства,
транспорта и архитектуры им. Н. Исанова
Учебно-издательский центр “Авангард”