МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Кафедра Систем Информатики
(название кафедры)

Р.,А., Джумамуратов
(И., О., фамилия студента – автора работы)

Разработка средств создания морфологических словарей казахского языка на основе корпуса размеченных текстов

(полное название темы магистерской диссертации)

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ
по направлению высшего профессионального образования

230100.68 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Тема диссертации утверждена распоряжением по НГУ № 90989 от «15» марта 2012г.
Тема диссертации скорректирована распоряжением по НГУ № 110449 от «18» декабря 2012г.

Руководитель
Сидорова Е.А.
(фамилия , И., О.)
к.ф.-м.н.
(уч.степень, уч.звание)

Новосибирск, 2013 г.

Оглавление

Введение

Существенную поддержку в проведении лингвистических исследований оказывают программные средства, позволяющие автоматически находить в исследуемых текстах нужные словоформы. Эту задачу отчасти решают специальные программы, которые выполняют поиск словосочетаний, используя сделанную заранее лингвистическую разметку текстов корпуса.

Важным, едва ли не центрообразующим звеном цепи автоматической обработки текста на естественном языке является технология нахождения основы слова (стемминг), родственный ей по целям алгоритм (лемматизация), позволяющий определить, что некоторая цепь словоформ составляет одно «словоизменительное гнездо» (имеет одну лемму). Конечным продуктом, способным на эти операции, является программа, в автоматическом режиме осуществляющая морфологический разбор слова[1].

Проблема обработки текстов на казахском языке, «понимания» языка компьютером была и остается актуальной. Среди множества задач, которые сводятся к решению данной проблемы, можно назвать такие, как общение с компьютером на естественном языке, информационный поиск, машинный перевод, извлечение содержательной информации из текстов, пополнение баз знаний и создание конкордансов – словарей, содержащих слова из всех работ одного автора. Достаточно рутинная работа – проанализировать стилистику какого - либо автора по его работам. Благодаря автоматическому разбиению слов на морфемы и статистическим данным, которые рассчитывает программа, появляется возможность автоматизированного анализа авторских текстов и составления готовых конкордансов.

В рамках работы было произведено изучение морфологии казахского языка.
Првильное понимание состава слова, умение определить образующие его компоненты имеют большое значение при изучении языка. В слове отражены особенности строя языка, его лексика - семантические и функционально - грамматические законы.

Казахский язык по своей типологии и морфологической структуре значительно шире, чем просто набор элементов лексики, и отличается относительной регулярностью, позиционной и грамматической стабильностью морфологической структуры различных словоформ. Образование слов происходит последовательного присоединения к основе слова грамматических частиц - аффиксов [2].

В целях построении модели морфологии казахского языка была проведена морфемно - морфологическая разметка (ММР) корпуса казахских текстов

Создание методов разметки казахского языка было реализовано в программе MarkSystem. Основное назначение, которой является выделение «значимых» фрагментов

текста, их сопоставление заданным категориям, поддержка функций редактирования, поиска, визуализации.

Создаваемый с помощью данной среды размеченный корпус текста может быть использован другими программными инструментами для автоматизирования создания различных лингвистических ресурсов. Другой немаловажной возможностью является «ручной» анализ экспертной разметки с целью выявления типичных понятий, фактов, отношений и их выражения в тексте. На основе анализа формируется система семантических категорий, описываются универсальные структуры (шаблоны) ситуаций.

1. Постановка задачи

Целью работы является разработка методов морфологического анализа текстов на казахском языке, а так же методов корпусного исследования текстов и создания предметных словарей.

Для разработки приложения были поставлены следующие задачи :

1. Изучение морфологии казахского языка, выделение морфологических классов, исследование структур парадигм.
2. Исследование существующих систем морфологического анализа текстов тюркских языков.
3. Построение морфологической таблицы для казахского языка.
4. Построение иерархии семантических признаков для разметки научных текстов
5. Создание семантической разметки корпуса научных текстов на русском и казахском языках.
6. Создание морфемно - морфологической разметки корпуса текстов на казахском языке на основе разработанной морфологической таблицы.
7. Разработать словарь аффиксов и начальных форм слов обеспечивающие эффективную обработку словоформы.
8. Разработать алгоритм морфологического анализа словоформ.
9. Разработать пользовательский интерфейс, позволяющий редактировать словарь основ, а так же проводить анализ словоформ.
10. Реализация программного модуля позволяющий производить морфологический анализ.

2 Представление структуры языка и систем анализа текста

2.1 Структурно - типологическая характеристика казахского языка

Структурно - типологическая характеристика казахского языка связана с его принадлежностью к агглютинативным языкам. Для описания языков агглютинативного типа применяется набор признаков, учитывающих не только морфологические, но и синтаксические и фонетические особенности. Морфологические признаки агглютинации: Сохранение постоянного фонетического облика корневой морфемы:

1.Корень слова (Түбір) – в именительном падеже выступает в чистом виде, таким образом является центром всей парадигмы склонения;

2. Между морфемами четко сохраняется граница;

3. Строгая последовательность присоединения аффиксов (жұрнақ + жалғау) (см. Рис. 1).

Основа	+ суффикс	+ мн. оконч.	+ оконч. принадл.	+ падежн. оконч.	+ личн. оконч.
ел	-ші-лік	-тер	- і	-нде	-
(біздің) қызмет	-кер	-лер	-іміз	-	-
ауыл	-	-дар	-ы	-нан	-быз

Рис.1. Определение присоеденения окончаний

Фонетические признаки агглютинации:

1. Наличие сингармонизма (үндестік заңы);

2. Фиксированное ударение (екпін), которое способствует сохранению фонетической целостности слова.

Синтаксические признаки агглютинации:

1. Твердый порядок слов в предложении:

a) определение (анықтауыш) находится перед определяемым словом;

b) Дополняющее слово (толықтауыш) находится перед дополня емым словом;

c) Сказуемое (баяндауыш) – в конце предложения [3].

2.2 Специфика казахского языка

Казахский язык имеет ряд специфических особенностей:

1. Ударение в казахском языке падает на последний слог слова. При прибавлении аффиксов (жұрнақ + жалғау) к слову ударение передвигается вместе с границей слова: мемле***кет*** – мемлекет***тер*** – мемлекеттер***ден***; әкім***дік*** – әкімдік***те*** – әкімдік***тен*** – әкімдіктен***біз***

2. Для обозначения принадлежности предмета в казахском языке употребляются не только притяжательные местоимения (тәуелдену), но и особые притяжательные окончания: менің әке - ***м*** - мой отец; сенің пәтер - ***ің*** – твоя квартира.

3. В казахском языке слова с личными окончаниями (жіктік жалғау) во многих случаях являются сказуемыми (баяндауыш). Эти окончания присоединяются не только к спрягаемым формам глагола (етістік), но и к другим словам, выступающим в предложении в качестве сказуемого: мен қызметкер - ***мін*** – я служащий

4. В казахском языке отсутствует категория рода. Поэтому вместо он, она, указывающих на человека, о котором идет речь, употребляется только одно местоимение – ол: он сказал – ол айтты она сказала – ол айтты

5. Одно и то же прилагательное (сын есім), местоимение (есімдік) или порядковое числительное (реттік сан есім), в зависимости от смысла предложения, может переводиться в разных родах: ақылды адам – умный человек; менің қалам – мой город; бесінші том – пятый том; ақылды қыз – умная девушка; менің елім – моя страна; бесінші тарау – пятая глава; ақылды сөз – умное слово; менің байлығым – мое богатство; бесінші терезе – пятое окно.

6. Предлоги большей частью передается посредством послелогов (септеуліктер) и в форме косвенных падежей; анама алдым – купил для матери тіл туралы – о языке жұмысқа дейін – до работы

7. расположение аффиксов в казахском языке всегда после корня только в одном направлении: ел – страна, ел - ***дер*** – страны, ел – ***дер - іміз*** – наши страны, ел – ***дер – іміз - ден***– от нашей страны.

Служебные слова (шылау сөздер), вспомогательные глаголы (көмекші етістіктер), служебные имена (көмекші есімдер), имеют такое расположение, послелоги, присоединяющиеся к глаголам и т.п.: ұмытып қала жаздап еді – чуть ли не забыл

8. Если перед существительным (зат есім) стоит числительное (сан есім), то окончание множественного числа (көптік жалғау) в существительном не употребляется:

жеті ***маман*** – семь ***специалист*** - ов

(жеті мамандар емес)

9. В казахском языке количественные (есептік), порядковые числительные (сан есім) при употреблении перед существительными в качестве определения, не изменяются в числе и падеже:

он кітап – ***десять*** книг

оныншы қатар – ***десятый*** ряд

оныншы адамға – ***десятому*** человеку [4].

2.3 Образование слов

Существуют закономерности в образовании и появлении новых слов. Несколько способов образования слов:

1. Например, слова оқушы – ученик, оқулық – учебник, аялдама – остановка образованы путем присоединения аффиксов - шы, - лық, - ма к основе, т.е., к корню слова: оқу – учить, аялда – остановись.

2. Путем изменения значения жеті күн – семь дней на апта, жеті – неделя. В процессе развития языка такие способы в образовании новых слов, получили научное название – словообразование (сөзжасам) [5].

2.3.1 Состав слова

Слово состоит из корня слова (сөз түбірі) и аффикса (қосымша).

Корень слова (сөз түбірі) – это неделимая часть слова, которая имеет самостоятельное лексическое значение: жол – дорога, жаз – пиши, мектеп – школа.

Аффикс (қосымша) состоит из: суффикса (жұрнақ) и окончания (жалғау).

Производные слова (туынды сөздер) – основа, состоит из корня и суффикса, называется производным словом: жол+дас – товарищ, жаз+у+ шы – писатель, бала+лық – детство, тәрбие – воспитание, тәрбие+ші – воспитатель.

Однокоренные слова (түбірлес сөздер) –образованы путем присоединения к одному корню различных аффиксов: өнім – продукция, өндіріс – производство.

Слово ***өн*** непроизводный корень.

Аффикс (қосымша) – суффикс (жұрнақ) и окончание (жалғау) находятся в слове после корня. Они подвергаются действию сингармонизма и прогрессивной ассимиляции.

2.3.2 Сложные слова

Сложные слова (күрделі сөздер) – образованы путем сложения, соединения двух корневых, повторения парных, а также сокращения сложных слов. Они являются

сложными видами корневых слов. По способу образования сложные слова (күрделі сөздер) делятся на: а) слитные слова (біріккен сөздер);

ә) парные слова (қос сөздер);

б) сокращенные слова, аббревиатура (қысқарған сөздер).

Слитные слова (біріккен сөздер) – это один из видов сложных слов, состоящих из сочетания двух корней, являющихся названием нового предмета или явления: алғысөз – предисловие, баспасөз – пресса, төлқұжат – пас - порт, жанармай – топливо.

Слитные слова бывают двух видов: В первом случае основа корневых слов сохраняется: тікұшақ – вертолет, саяжай – дача. Во втором случае, вошедшие в состав слитных слов корневые слова подвергаются звуковым изменениям (кіріккен сөздер): жаздыгүні (жаздың күні) – летом, түрегелді (тұра келді) – встал, поднялся, – таким образом и т.д. Слитные слова составляют следующие названия:

1. Имена людей: Нұрсұлтан, Қасымжомарт, Бауыржан, Дастан, Әуез, т.д.

2. Анатомические названия: асқазан – желудок, тоқішек – толстая кишка, соқырішек – аппендицит и т.д.

3. Географические названия (местности, планеты, звезд, и т.д.): Қызылорда, Ақтау, Сырдария, Алакөл, Жетіқарақшы – Большая медведица;

4. Названия песен, кюев: Саржайлау, Сарыарқа, Қараторғай, Кісенашқан, Сегізаяқ;

5. Другие названия: шекара – граница, кәсіпорын – производство.

Слитные слова (біріккен сөздер) пишутся вместе.

2.4 Специализированный корпус

Корпус — это информационно - справочная система, основанная на собрании текстов на некотором языке в электронной форме. Специализированный корпус содержат тексты определенного типа при создании такого корпуса текста производится лингвистическое аннотирование (морфологическое, синтаксическое), не зависящее от ПО и осуществляемое автоматически и/или вручную.. Применяется два вида аннотирования: терминологическая разметка которая фиксирует присутствие в тексте понятий ПО, разметка отношений(в частности ситуационная разметка).

Разметка — главная характеристика корпуса; она отличает корпус от простых коллекций (или «библиотек») текстов, в изобилии представленных в современном интернете. От степени разнообразия разметка, зависит научная и учебная ценность корпуса.

Размеченные фрагменты текста используются для наполнения предметного словаря. Отмеченная лексика обрабатывается морфологическим и синтаксическим компонентами

словарной технологии, нормализуется, вносится в словарь и снабжается семантическими признаками в соответствии с разметкой.

Ситуационная разметка планируется использовать для анализа контекстов предикатных лексем с целью автоматизированного наполнения словаря семантико - синтаксических шаблонов [6].

Раньше исследователь мог лишь просматривать тексты и вручную выписывать из них нужные примеры; эта предварительная (но абсолютно неизбежная) деятельность была очень трудоемкой и не позволяла обрабатывать большие массивы материала. Теперь ограничений на объем анализируемого материала и скорость поиска информации в нем по существу нет, а это означает, что в распоряжении исследователя оказываются колоссальные массивы текстов самого разного типа. Это не замедлило сказаться на развитии наших знаний о языке: возможность массовой — в том числе статистической — обработки текстов, недоступная прежде, позволила обнаружить в структуре и развитии языка такие закономерности, о существовании которых наука раньше или не подозревала, или лишь смутно догадывалась, но не могла строго обосновать. Теперь подлинно научные описания грамматического строя языков, а также авторитетные академические словари — практически все без исключений — должны составляться на основе корпусов этих языков. Учет корпусных данных оказывается крайне желательным (если не строго обязательным) и при многих других более специальных научных исследованиях [7].

2.4.1 Синтаксическое аннотирование

Глубоко аннотированный (синтаксический) корпус - данный фрагмент Национального корпуса содержит тексты, снабженные морфосинтаксической разметкой. Это значит, что помимо морфологической информации, приписанной каждому слову текста, для каждого предложения задана его синтаксическая структура.

Синтаксическая структура предложения, используемая в глубоко аннотированном корпусе (ГАК), представляет собой дерево зависимостей, в узлах которого стоят слова предложения, а ветви помечены именами синтаксических отношений. Такое представление о синтаксической структуре предложения восходит к лингвистической модели «Смысл ⇔ Текст» И.А.Мельчука и А.К.Жолковского. Окончательный перечень синтаксических отношений, используемых в ГАК, а также целый ряд конкретных лингвистических решений, связанных с представлением синтаксической структуры предложения.

2.5 Обзор методов анализа текстов

Самые большие возможности и высокое качество анализа текстов можно получить, проведя его полный анализ. Для полноценной работы анализа текста нужно проанализировать тест, с точки зрения синтаксиса (структуры предложений), семантики (понятий, применяемых в тексте) и прагматики (правильности употребления понятий и целей их употребления). В целом для проведения полного анализа необходимо создать следующие методы:

Графематический анализ – обеспечивает выделение синтаксических или структурных единиц из входного текста, который может представлять собой линейную структуру, содержащую единый фрагмент текста. В более общем случае текст может состоять из многих структурных единиц: основного текста, заголовков, вставок, врезок, комментариев и т.д. Графематический анализ должен выделять синтаксические единицы: абзацы, предложения, отдельные слова и знаки препинания. В ряде случаев здесь же проводится предморфологический анализ – объединение неразрывных неизменяемых словосочетаний в одну единицу.

Морфологический анализ – обеспечивает определение нормальной формы, от которой была образована данная словоформа, и набора параметров, приписанных данной словоформе. Это делается для того, чтобы ориентироваться в дальнейшем только на нормальную форму, а не на все словоформы, использовать параметры, например, для проверки согласования слов. Морфологическая структура словоформы представляет собой имя лексемы, или лемму, которой приписывается часть речи и морфологические характеристики, т.е. значения соответствующих морфологических категорий.

Синтаксический анализ – самая сложная часть анализа текста. Здесь необходимо определить роли слов и их связи между собой. Результатом этого этапа является набор деревьев, показывающих такие связи. Выполнение задачи осложняется огромным количеством альтернативных вариантов, возникающих в ходе разбора, связанных как с многозначностью входных данных (одна и та же словоформа может быть получена от различных нормальных форм), так и неоднозначностью самих правил разбора.

Семантический анализ проводит анализ текста «по смыслу». С одной стороны, семантический анализ уточняет связи, которые не смог уточнить постсинтаксический анализ, так как многие роли выражаются не только при помощи средств языка, но и с учетом значения слова. С другой стороны, семантический анализ позволяет отфильтровать некоторые значения слов или даже целые варианты разбора как «семантически несвязные». При такой разметке большинству слов в тексте приписывается один или несколько семантических и словообразовательных признаков. Например:

'вещество', 'пространство', 'скорость', 'движение', 'обладание', 'свойство человека', 'отглагольное имя' и т.п. [8].

2.6 Обзор систем морфологического анализа текста

В рамках данной работы были рассмотрены различные проекты, позволяющие производить морфологический анализ текста это интеллектуальный морфологический анализатор основанный на семантический сетях и морфологический анализатор башкирского языка «bashmorph»

2.6.1 Интеллектуальный морфологический анализатор

Для формализации правил добавления суффиксов и окончаний предлагается использовать семантическую нейронную сеть. С помощью такой сети генерируются словоформы казахского языка, и порождается структура словаря начальных форм в виде синхронизированного линейного дерева. Для представления словоформы и ее признаков используются следующие метасимволы:

- разделитель между словами,

(- начало слова,

) - конец слова,

! - начало признака словоформы (падеж и т.д.),

* -конец признака словоформы.

пример слова «бала - ребенок» (основа слова) и двух его словоформ «балам - мой ребенок», «балаң - твой ребенок» (в казахском языке одушевленные существительные изменяются по лицам с помощью личных окончаний). Рецептор возбуждается на символ начала слова «(». Далее переходит в состояние «б», при подаче символа «б», далее последовательно «(ба», «(бал», «(бала» , и затем одновременно два субсостояния «(балам)» и «(балаң)»

структура связей леммы определяет следующие признаки: имя существительное (зат есім) –«!зе*» , одушевленное – «!жа*», притяжательное окончание (тәуелдік жалғау) первого лица – «!11*» (бірінші жақ),притяжательное окончание (тәуелдік жалғау)второго лица – «!22*» (екінші жақ). При подаче на лемму слова «(балам)» она переходит в возбужденные субсостояния: «(балам)», «!зе*», «!жа*», «!11*» а при подаче слова «балаң» в возбужденные субсостояния: «(балаң)», «!зе*», «!жа*», «!22*».

Нейроны распознают отдельные символы входной символьной последовательности. На выходе генерируется сигнал, означающий наличие или отсутствие соответствующего символа в анализируемом тексте. Нейроны выдают результат распознавания отдельных фрагментов входной символьной последовательности. Для обозначения таких фрагментов во входной символьной последовательности применяются метасимволы скобок: "(" и ")".

Тогда приведенный пример перепишется в виде: ((бала)м), ((бала)ң), (((бала)м)ның), (((бала)ң)да) [9].

2.6.2 Морфологический анализатор башкирского языка «bashmorph»

Морфологический анализатор башкирского языка создан в лаборатории в сентябре 2012 года. Программа «Basmorph» предназначена для разбора словоформ башкирского языка, установления их основы, состава и грамматического значения аффиксов, добавляемых к основе при словоизменении и отчасти словообразовании (программа умеет определять словообразовательный аффикс абстрактных существительных *-лыҡ/-лек* и аффикс деятеля *-сы/-се*). Воспользоваться анализатором как онлайн - сервисом можно на странице http://lcph.bashedu.ru/cgi-bin/parser.pl

Разбор башкирских форм представлен в четырёх равнозначных вариантах: на русском, башкирском, английском языках и в виде стандартного вывода программы, где граммемы даются в виде сокращённых обозначений, по возможности соответствующих Лейпцигским правилам глоссирования.

Грамматические правила, заложенные в логику парсера, основаны на академических описаниях башкирской грамматики и дополнены неучтёнными в грамматиках наблюдениями над реальным функционированием языка.

Вывод сформирован по образцу русского парсера Mystem. Однако у башкирского анализатора есть свои особенности. В частности, добавлена возможность представления русскоязычных эквивалентов значений найденных основ. Эта возможность пока охватывает не весь состав словника, внутренний словарь программы находится в стадии пополнения [10].

2.6.3 Сравнение систем морфологического анализа текста

Первая из рассмотренных систем – интеллектуальный морфологический анализатор – представляет собой гибкий инструментарий для обработки текста. По мимо анализа текста происходит наполнения словаря которые могут применятся в орфографии. Но данная система не предоставляет создания терминологических словарей.

Есть возможность использования полученных формализаций, методов и алгоритмов в системах обработки естественно-языковых текстов (орфографических корректорах, переводчиках, обучающих системах) и т.д.

Вторая расмотренная система - «bashmorph» - представляет инструментарий для анализа башкирского языка. Форма слова визуально предоставлена в различных вариантах языка. Появляется возможность создания автоматического переводчика с башкирского на

русский и английский языки и обратно. С помощью «bashmorph» можно создавать частотные словари, которые включают лингвистические единицы (словоформы, словосочетания), которые в ходе исследования текста регистрируются составителем. И указывается частота употребления в данном тексте. А так же заниматься исследованием лексической и грамматической структуры башкирских текстов, ставить промышленные задачи информационного поиска.

В процессе изучения казахской морфологии возникла необходимость создания своего инструментария, который можно было внедрить в систему извлечения терминов казахского языка. Для извлечения терминов необходим модуль морфологического анализа, который бы работал на уровне отдельных слов и приводил слово и его атрибуты в морфологическую норму.

3 *Корпусное исследование подъязыка предметной области*

В казахском языке концепция слова значительно шире, чем просто набор элементов лексики, и отличается относительной регулярностью, позиционной и грамматической стабильностью морфологической структуры различных словоформ. Слова в нем образуются присоединением к корню или основе слова грамматических частиц – аффиксов [11].

В целях построении модели морфологии казахского языка была проведена морфемно - морфологическая разметка (ММР) корпуса текстов (см. Рис. 2).

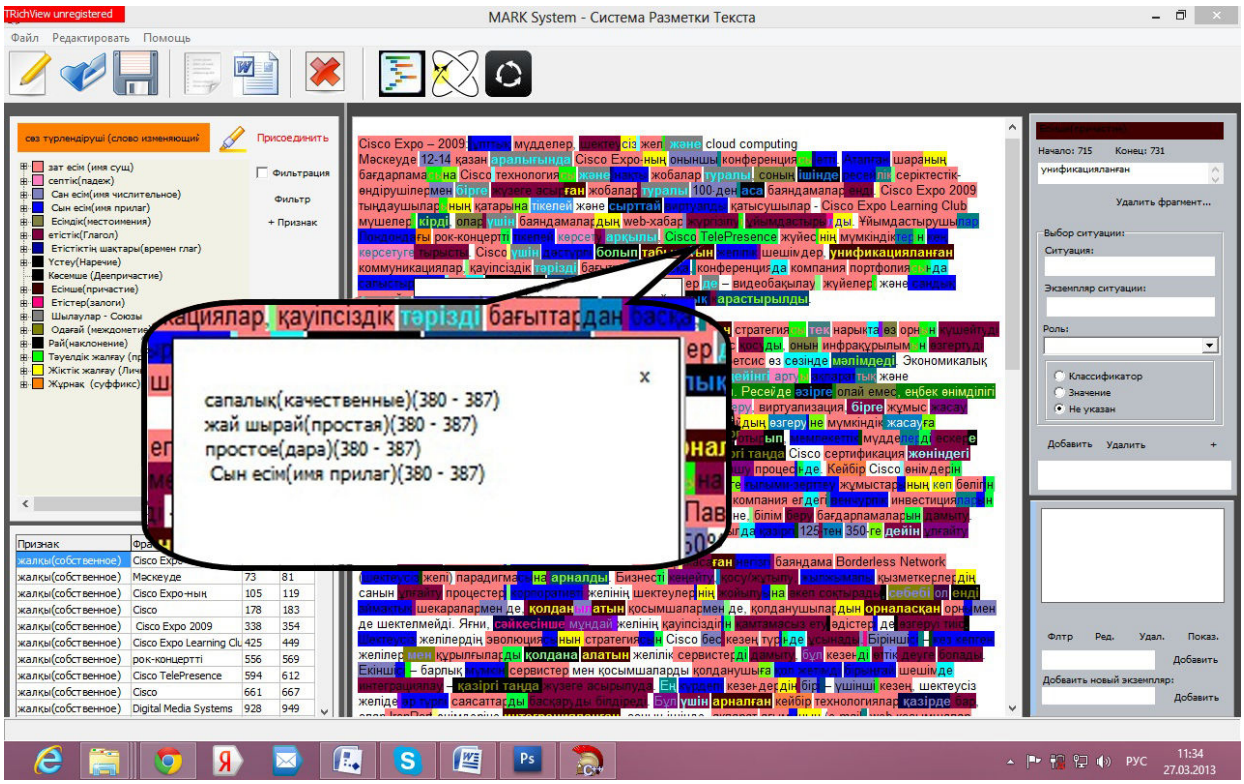

Рис. 2. Морфемно-морфологическая разметка казахского текста.

Разметка была создана с помощью системы MarkSystem [12]. Помимо основного функционала, связанного с непосредственным просмотром текста и его «раскраской», в системе реализуется возможность динамического создания и пополнения системы признаков (признаками помечаются фрагменты текста) и отношений, редактирования цветовой схемы «раскраски» (посредством сопоставляется признакам цветовой схемы) и управление визуализацией (фильтрация, конкорданс). Каждому фрагменту может быть сопоставлено множествово признаков и связей.

Система MarkSystem ориентирована как на лингвистов, осуществляющих лингвистическое исследование корпуса текстов, так и на экспертов, отмечающих

терминологию, характерную для заданной предметной области. Параллельно ММР, на другом уровне, создавалась семантическая разметка (СР) тех же самых текстов. Семантическая разметка текста ориентирована на заданную предметную область и включает терминологическую разметку и разметку тезаурусных отношений.

В процессе создания ММР и СР были разработаны иерархии признаков (см. Рис. 3).

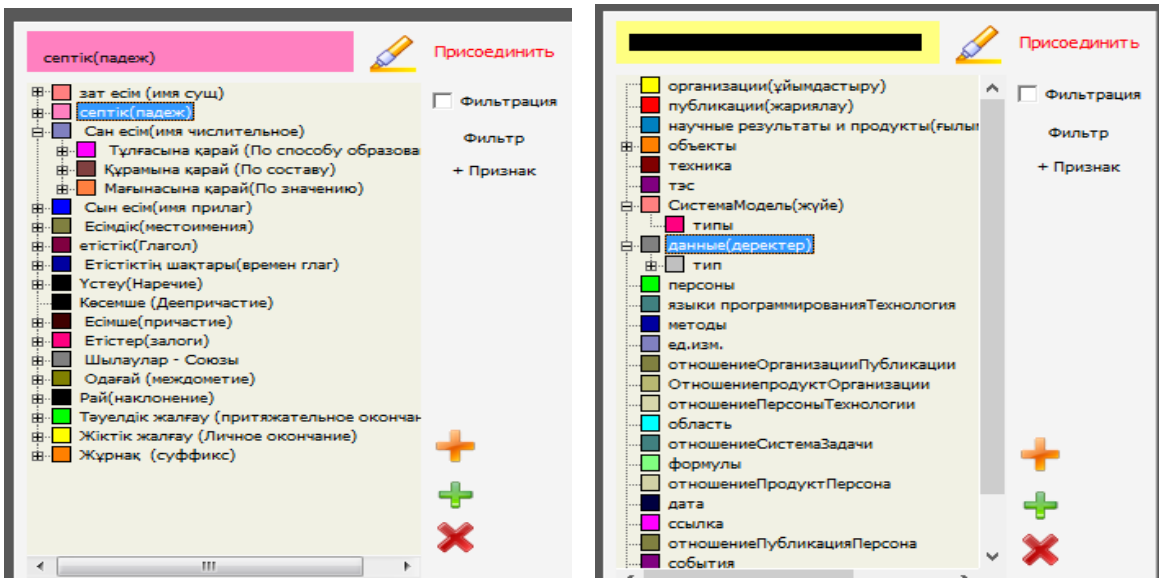

Рис. 3. Древо признаков морфемно-морфологической и терминологической разметки.

На основе ММР была создана морфологическая таблица, описывающая модель казахского языка. Иерархия признаков ММР содержит 117 вершин, на основе которых было сформировано 17 морфологических атрибутов (в том числе и часть речи) и выделено 36 морфологических классов терминов-лексем. По типу словоизменения классы сгруппированы в группы, для каждой из которых формруется свой список окончаний и функция, сопоставляющая окончанию значение одного из атрибутов. Данная таблица легла в основу создаваемого морфологического словаря казахского языка.

Иерархия признаков СР соответствует верхнему уровню иерархии терминов тезауруса по информатике и на текущий момент включает 29 вершин. В процессе СР любой термин, выделенный экспертом, либо соотносится с признаком (что означает, что эксперт считает данный термин обобщающим классом предметной области), либо связывается с одним из ранее введенных признаков. Выделенные фрагменты также могут связываться специальным отношением Род-вид (подробнее об этом будет сказано ниже). Таким образом, назначение СР – выделить основные классы терминов предметной области, которые впоследствии будут служить основой автоматизированных методов поиска новых терминов и отношений между терминами.

3.1 Морфологический анализ казахского текста

На вход морфологического анализатора подается упорядоченный список словоформ (с учетом знаков), полученного в результате графематического анализа. Для каждой словоформы на первом этапе осуществляется нормализация, т.е. поиск основы – начальной формы слова. Затем, в зависимости от части речи и найденных аффиксов, вычисляются морфологические характеристики слова.

3.1.1 Алгоритм морфологического анализа

Алгоритм морфологического анализа по правилу в тексте заключается в следующем:

Модуль нормализации в процессе своей работы осуществляет следующую последовательность шагов:

1 шаг: Выполняется поиск слова в словаре начальных форм. Если слово в словаре найдено, то шаг 5.

2 шаг: Слово считывается посимвольно в обратном порядке (начиная с конца слова). Если слово закончилось, то работа алгоритма завершается. На основе текущего списка аффиксов формируется список гипотетических аффиксов.

3 шаг: Выполняется поиск всех гипотетический аффиксов в словаре аффиксов. Все найденные аффиксы добавляются в список аффиксов. Если ни один новый аффикс не найден, то переходим к шагу 2.

4 шаг: Выполняется поиск начальной части слова в словаре начальных форм. Если слово не найдено, то переходим к шагу 2.

5 шаг: В результат добавляется найденная основа и сопутствующий набор аффиксов. Переход к шагу 2.

3.2 Определение нормальной формы слова

После нормализации, для каждого найденного слова осуществляется вычисление его морфологических характеристик на основе его аффиксов и морфологического класса основы.

Продемонстрируем результат морфологического анализа на примере (Рис. 4).

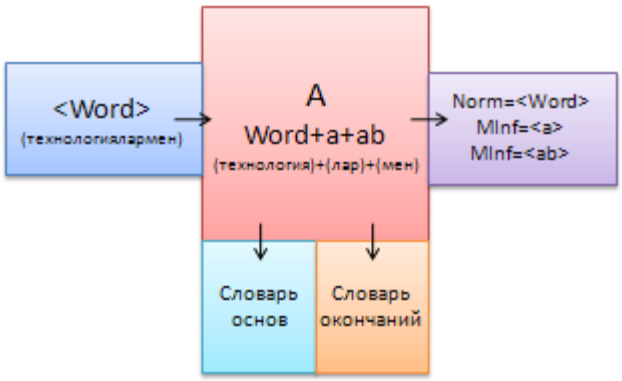

Рис. 4. Процесс определения нормальной формы слова и его морфологических параметров.

На вход подается словоформа *технологиялармен*, происходит поиск в словарях аффиксов *мен, лар* и основы *технология*. На основе морфологического класса основы (сущ.) и аффиксов вычисляем морфологическою информацию: *лар* <мн.число>, *мен* <род.падеж>.

Казахский язык характеризуется строгой последовательностью присоединения аффиксов к корню: вначале присоединяется словообразовательный суффикс, потом словоизменяющие окончание: принадлежности, падежей, лица и числа. Для имен существительных к основе слова вначале добавляется окончание множественного числа затем притяжательное окончание, далее следует падежное окончание и последним – окончание формы спряжения [2].

3.3 Правило присоединения окончаний в казахском языке

Окончания в казахском языке прибавляются по определенному правилу, которое представлено в таком виде:

$$С=ОС+КЖ+ТЖ+СЖ+ЖЖ, \quad (1)$$

где С – словоформа; ОС – основа слова; КЖ - окончание множественного числа; ТЖ - притяжательное окончание; СЖ - падежное окончание; ЖЖ - окончание формы спряжения.

3.3.1 Процесс образования нормальной формы слова

Морфологический анализатор должен определять по словоформе нормальную форму слова.

Нормальная форма слова – это форма слова (строка), принятая для обозначения понятия, связанного с данным словом Словоформа – это форма слова (строка),

связанная с нормальной формой слова и указывающая на особенности употребления данного слова. Будем считать, что <wform> характеризуется пятеркой – <line Wforms; PSpeech; Nform; PSpeech Nform; MorphParam>.

Wform - словоформа;

Line Wform - строка словоформы

PSpeech - часть речи

Nform - нормальная форма, от которой была образована данная словоформа;

PSpeech - часть речи нормальной формы;

Nform - нормальной формой, от которой была образована данная словоформа

MorphParam - набор морфологических параметров, приписываемых к данной словоформе.

4 Программная реализация морфологического анализатора

4.1 Структура разработанной системы

В процессе извлечения терминов из документа исходный текст подвергается графематическому (разбиение на слова), морфологическому (определение нормальной формы и набора параметров) и поверхностно-синтаксическому (сборка словосочетаний) анализу (Рис. 5).

Рис 5. Общая схема выделения терминов

На этапе графематического анализа, после разбиения текста на слова, происходит поиск составных слов, которые должны рассматриваться как одно (с точки зрения морфологического анализатора). Морфологический анализ работает на уровне отдельных слов (в том числе составных) и возвращает морфологическую норму и атрибуты данного слова. При этом может оказаться, что одной словоформе может быть сопоставлено несколько возможных вариантов слов. Синтаксический анализатор может осуществлять поиск словосочетаний на основе синтаксичеких шаблонов сборки именных групп, аналогично [13-14]. В результате анализа приведенные к нормальному виду слова и словосочетания помещаются в предварительный словарь терминов.

На текущий момент недоступны программные инструменты, проводящие морфологический анализ текстов на казахском языке. Поэтому нами был разработан специализированный модуль и морфологическая модель казахского языка для системы Klan [14], предназначенной для автоматизированного создания терминологических словарей. Эта же система использовалась для морфологического синтаксического анализа текстов на русском языке.

Поиск терминов-словосочетаний осуществляется на основе правил разработанных в рамках системы Klan для русского языка и спроецированных на морфологическую таблицу казахского языка. Учет дополнительных особенностей языка в плане образования

устойчивых словосочетаний требует привлечения казахских специалистов и является одной из ближайших целей проекта.

В рамках данной работы было создано программное приложение, позволяющее проводить анализ слов казахского языка, работать со словарем начальных форм слова (редактировать, удалять, добавлять новые формы слов), словарем окончаний. А также осуществлять обработку слова словарями начальных форм и окончаниями слова, правилами морфологического анализа, т.е. определение основы формы, от которой была образована данная словоформа, и набора параметров, приписанных данной словоформе.

Данный модуль является компонентной программного комплекса обеспечивающего выделение терминов из текста. Система состоит из 4 модулей: графематического анализа, морфологического анализа, синтаксического и генератора словарей (см. пример на Рис.5).

Анализатор словоформ позволяет (см. пример Рис. 6.):

- Производить наполнение словаря начальных форм, если в словаре нет такой формы.
- При наполнении словаря у пользователя есть возможность удаления словоформы, если это не корректная форма слова.
- Модуль представляет возможность просмотра результа обработки слова.
- Отображается найденная в словаре основа слова, аффиксы которые были извлечены из словоформы.
- Есть возможность загрузки словарей основ и аффиксов

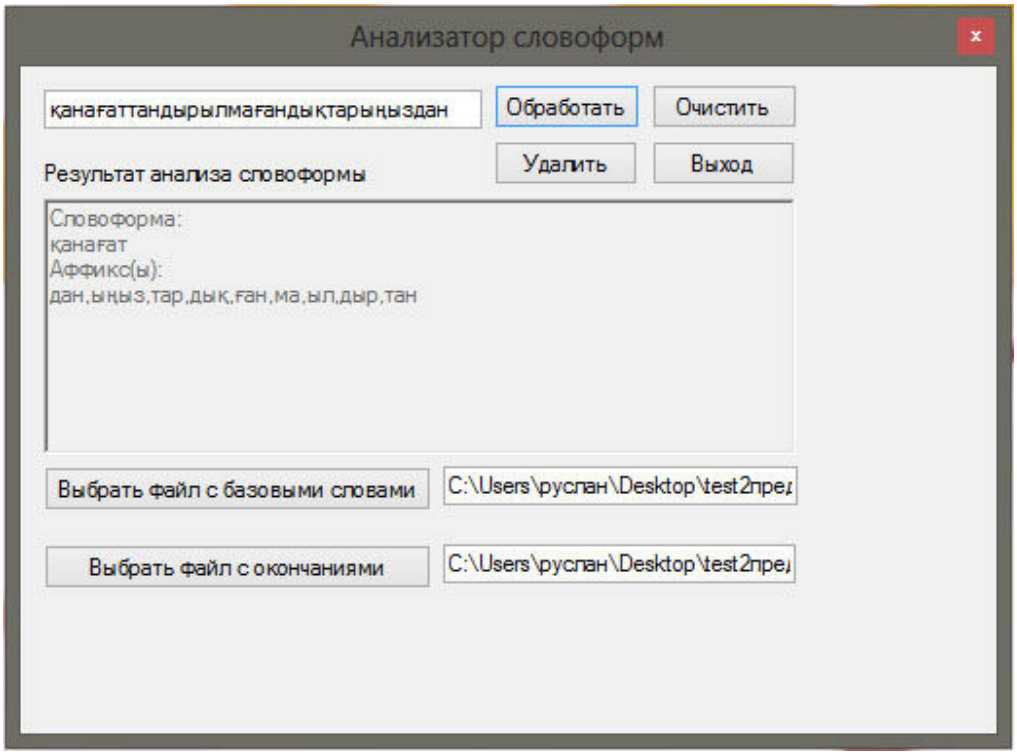

Рис.6. Интерфейс анализатора казахских словоформ

В текстовое поле вводится слово для обработки при нажатии на кнопку обработать, начинается работа алгоритма по завершению, если такого слова нет в словаре, то происходит автоматическое добавление слова в словарь.

В противном случае в результате отображается найденная основа и список извлеченных аффиксов.

4.2 Проведенные тесты

Таблица 1. Результаты слов после обработки

Название словоформы	аффиксы	Колличество аффиксов
Қанағат	[тан] [дыр] [ыл] [ма] [ған] [дық] [тар] [ыңыз] [дан]	9
Қам	[сыз] [дан] [дыр] [ыл] [ма] [ған] [дық] [тан]	8

Было протестированны слова с разными вариациями суффиксов и окончаний (см.таб. 1).

В таблице представлены результаты протестированых слов с наибольшим колличеством аффиксов.

Из полученных данных видно, что найдены все аффиксы и сама форма слова.

Заключение

В процессе решения поставленных в работе задач были достигнуты следующие цели:

- Изучение морфологии казахского языка, выделение морфологических классов, исследование структур парадигм.
- Построение морфологической таблицы для казахского языка
- Построение иерархии семантических признаков для разметки научных текстов
- Создание семантической разметки корпуса научных текстов на русском и казахском языках
- Создание морфологической разметки корпуса текстов на казахском языке на основе разработанной морфологической таблицы.
- Разработано представление словоформ
- Реализован алгоритм анализа слов
- Создана визуальная оболочка, позволяющая производить анализ словоформ редактировать, наполнять словарь новыми основами.
- Проведены тесты

Данное приложение планируется внедрить в систему выделения терминов в качестве модуля. В дальнейшем планируется расширить возможности пользовательского интерфейса, расширить процесс анализа словоформ, добавить определения частей речи слова и параметры аффиксов.

Литература

1. Орехов Б. В. Слободян Е. А. Проблемы автоматической морфологии агглютинативных языков и парсер башкирского языка [Текст] / // Информационные технологии и письменное наследие: материалы международной научной конференции (Уфа, 28–31 октября 2010 г.) / отв. ред. В. А. Баранов. — Уфа; Ижевск: Вагант, 2010. — С. 167–171.
2. Сопоставительная грамматика русского и казахского языков. Морфология. – Алма - Ата: Наука, 1966. – 459 с.
3. Балакаев М.Б. Современный казахский язык, (Фонетика и морфология). – Алма – Ата: Издательство Академии наук КазССР, 1962. — С. 451
4. Бектурова А., Бектуров Ш. Казахский язык для всех – Алматы: Ата - мұра, 2004. — С. 720
5. Франк В. Казахский язык (пособие для старших классов средних школ и студентов вузов)./ отв. Ред. Франк В.) - Астана: ТОО «АкПол», 2003. — С. 135
6. Кононенко И. С., Сидорова Е. А. Система семантической разметки корпуса текстов как инструмент извлечения экспертных знаний (на материале текстов по катализу) // Труды международной конференции «Корпусная лингвистика — 2011». — Санкт-Петербург, 2011. — С. 193–198.
7. Что такое корпус? Электронный ресурс. – Режим доступа: http://www.ruscorpora.ru/corpora-intro.html,
8. Клышинский Э.С. Начальные этапы анализа текста — М.: Издательство МИЭМ, 2011. — С.272
9. Шарипбаев А.А., Беманова Г. Т. Построение логической семантики слов казахского языка Евразийский национальный университет им. Л. Н. Гумилева, ул. Мунайтпасова 5, г. Астана, 010000, Казахстан. Материалы Всероссийской конференции с международным участием «Знания–Онтологии-Теории» (ЗОНТ-2011) 3-5 октября 2011 г., Новосибирск, 2011. – С. 156 – 163.
10. Орехов Б., Галлямов А. Башкирский морфологический анализатор. Электронный ресурс – режим доступа: http://lcph.bashedu.ru/index.php?go
11. Сидорова Е.А. Многоцелевая словарная подсистема извлечения предметной лексики // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2008». Вып. 7 (14). –М.: РГГУ, 2008. – С. 475-481.
12. Сидорова Е.А., Загорулько М.Ю. Программный инструментарий разработки лингвистических ресурсов // Труды III Международной научно-технической конференции

«Открытые семантические технологии проектирования интеллектуальных систем» OSTIS-2013. –Минск: БГУИР, 2013. –С.159-164.

13. Большаков И.А. Какие словосочетания следует хранить в словарях? // Труды международного семинара Диалог’2002 по компьютерной лингвистике и ее приложениям. Протвино: 2002. Т.2. С.61–69.

14. Загорулько М.Ю., Сидорова Е.А.. Система извлечения предметной терминологии из текста на основе лексико-синтаксических шаблонов // Труды XIII Международной конференции «Проблемы управления и моделирования в сложных системах» / Под ред.: акад. Е.А. Федосова, акад. Н.А. Кузнецова, проф. В.А. Виттиха. – Самара:Самарский научный центр РАН, 2011. – С.506-511.