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

Джумамуратов Р.А.

Группа 7205

Научный руководитель

канд. физ.-мат. наук Сидорова Е.А.

Актуальность работы

- Проблема развития средств морфологического анализа текстов на казахском языке.

- Задачи, которые сводятся к решению данной проблемы, извлечение содержательной информации из текстов, пополнение баз знаний и создание конкордансов – словарей.

Цель работы

- Разработка методов морфологического анализа текстов на казахском языке, а так же методов корпусного исследования текстов и создания предметных словарей.

Задачи работы

- Изучение морфологии казахского языка, выделение морфологических классов, исследование структур парадигм.
- Исследование существующих систем морфологического анализа текстов тюркских языков.
- 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 морфемно - морфологической разметки корпуса текстов на казахском языке на основе разработанной морфологической таблицы.
- Разработать словарь аффиксов и начальных форм слов обеспечивающие эффективную обработку словоформы.
- Разработать алгоритм морфологического анализа словоформ.
  Реализация программного модуля позволяющий производить морфологический анализ.

Обзор существующих систем

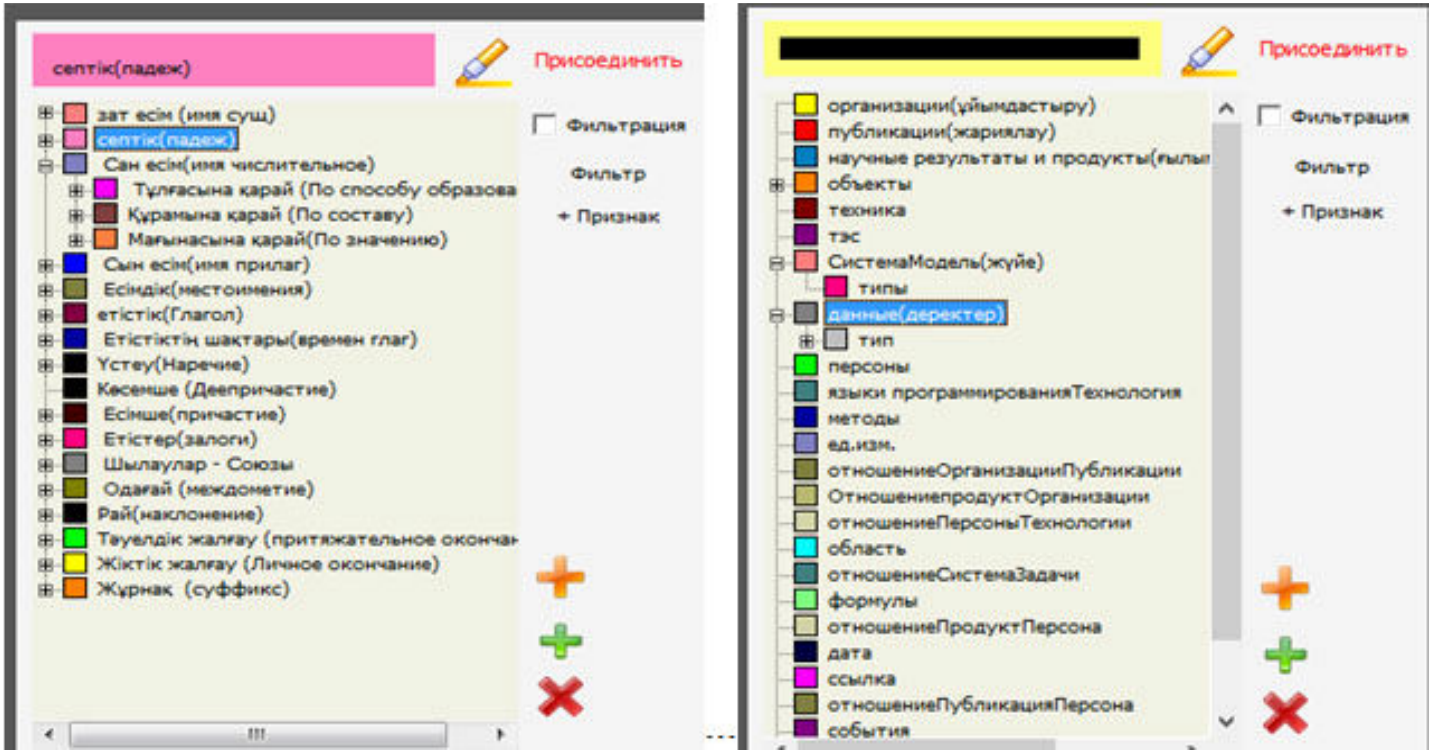
- Интеллектуальный морфологический анализатор казахского языка
- Морфологический анализатор башкирского языка «bashmorph».

Правило присоединения окончаний

$$С=ОС+КЖ+ТЖ+СЖ+ЖЖ$$

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

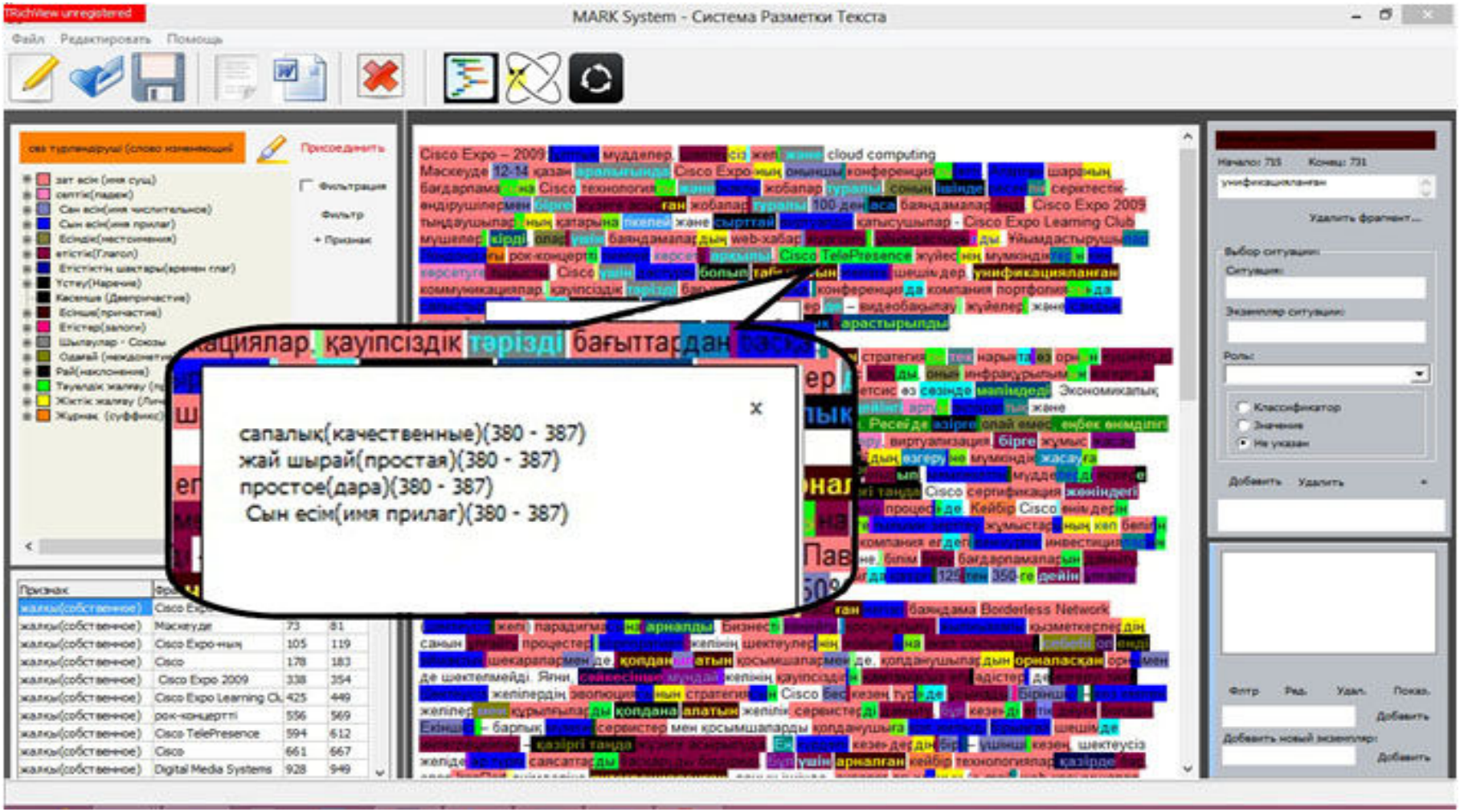
Иерархия признаков



Количество признаков для морфологической разметки - 17

Количество признаков для семантической разметки – 25

Морфемно - морфологическая разметка



Для ММР были подобраны корпусы текстов на русском и казахском языках объемом 50 - 100 статей (колличество символов 200 000), жанр – техническая литература.

Алгоритм морфологического анализа

- **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.

Определение нормальной формы слова

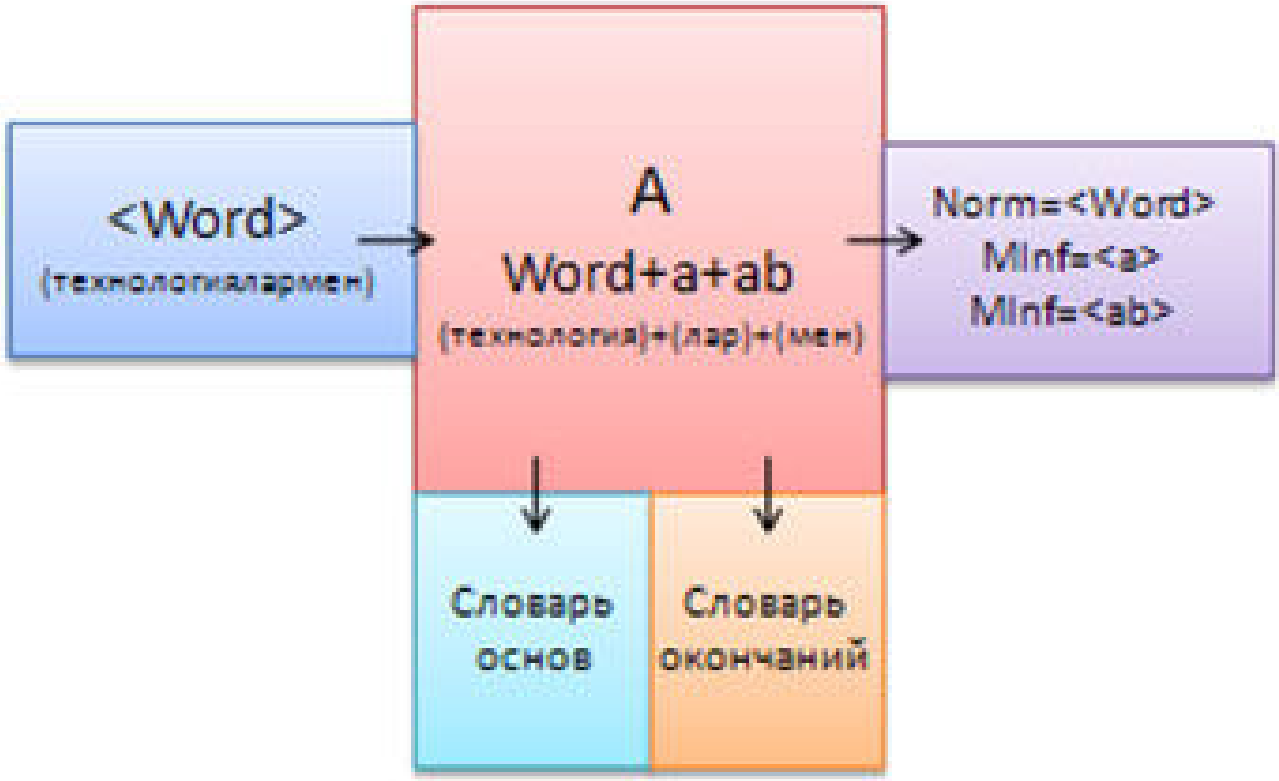
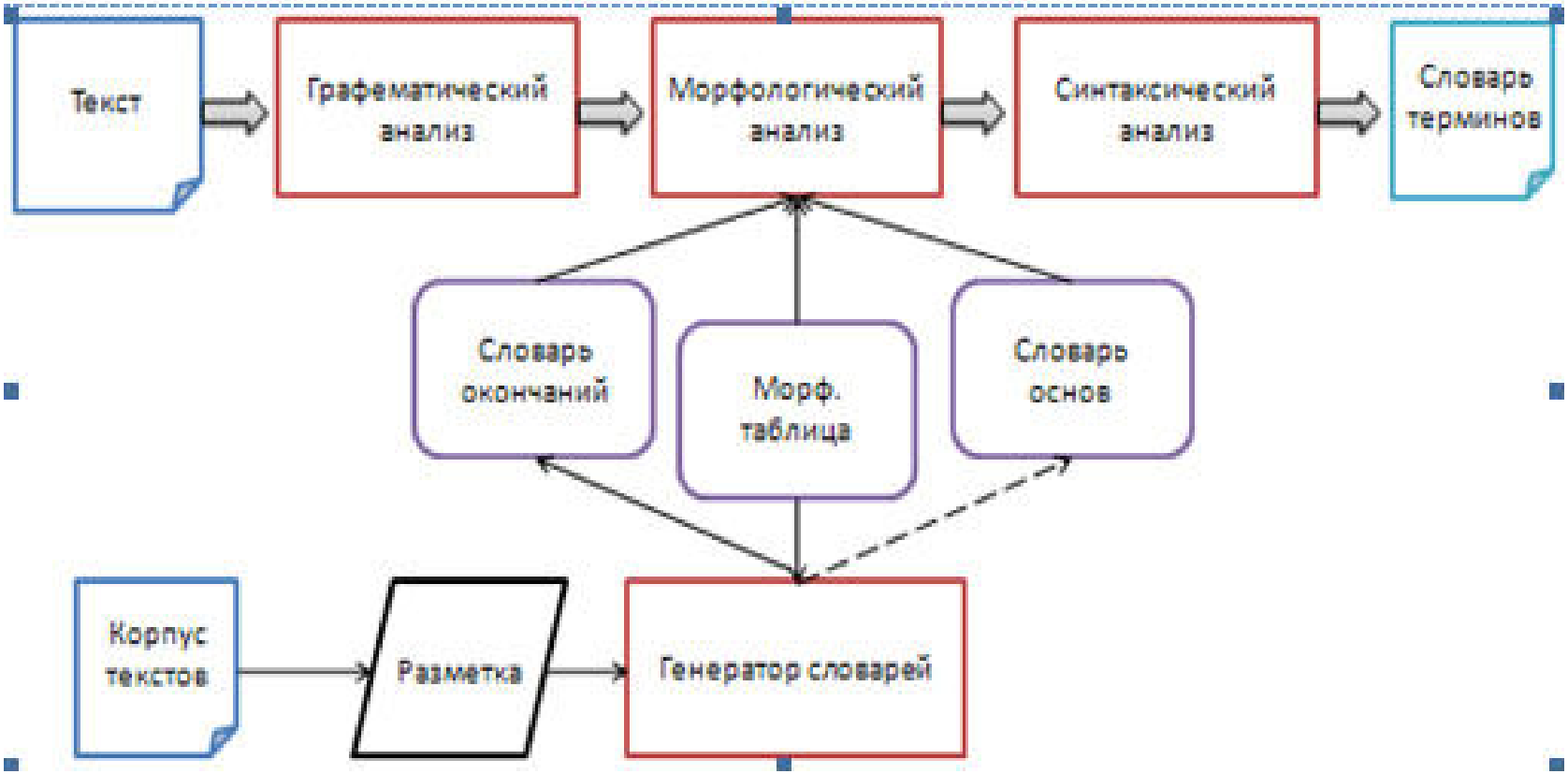
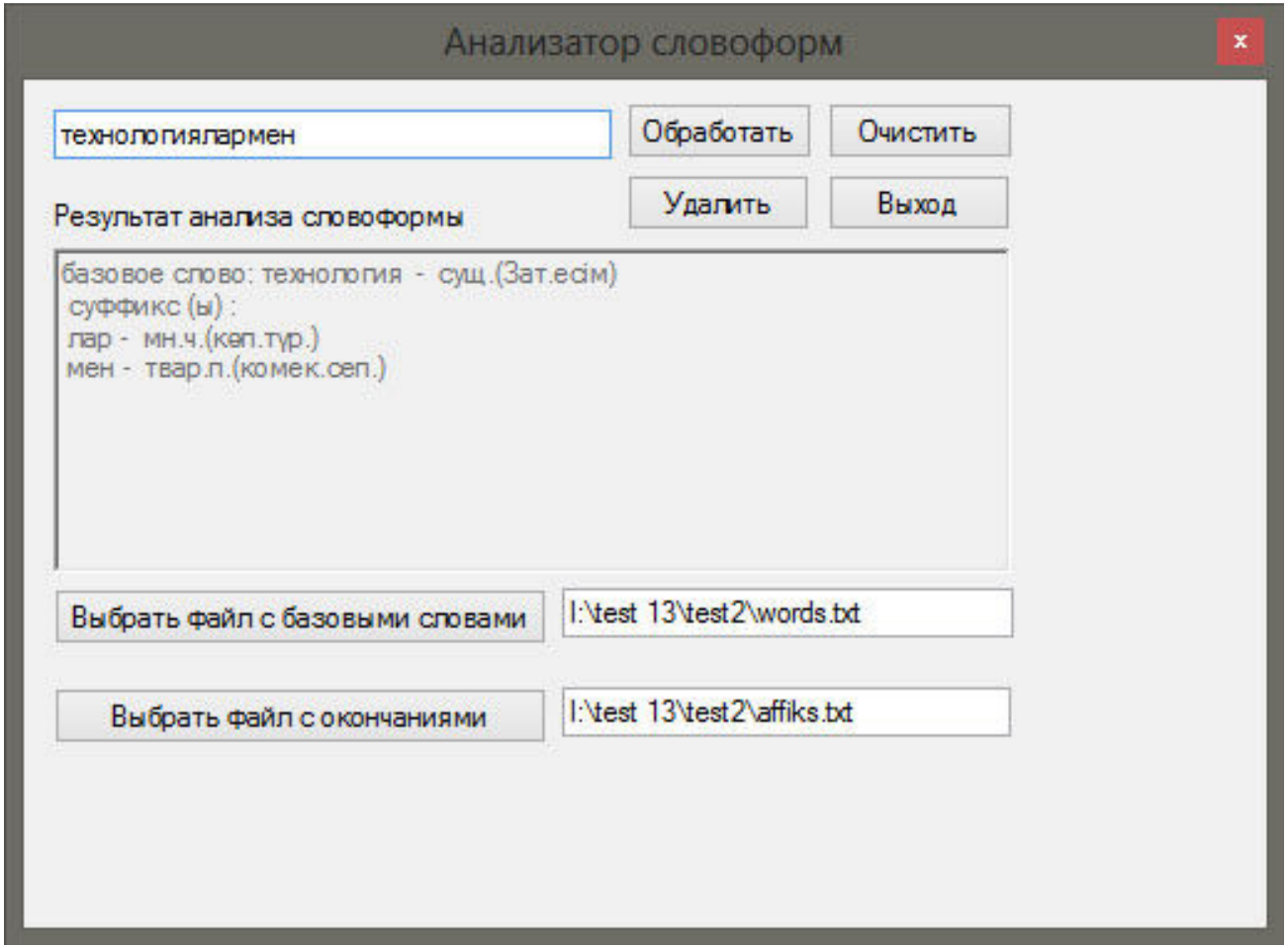


Схема выделения терминов



Пример обработки словоформы



Проведенные тесты

- 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

Заключение

- Построена морфологическая таблица языка и изучено представление слов казахского языка
- Построена иерархия семантических признаков для разметки научных текстов
- Создана семантическая разметка корпуса научных текстов на русском и казахском языках
- Создана морфологическая разметка корпуса текстов на казахском языке на основе разработанной морфологической таблицы.
- Реализован алгоритм анализа слов
- Создана визуальная оболочка, позволяющая производить анализ словоформ редактировать, наполнять словарь новыми основами.

Публикации

- Материалы 50-й юбилейной международной научно-технической конференции «Студент и научно-технический прогресс», НГУ, 2012. Информационные технологии.
- Вестник БГУ, выпуск 9, 2013

Спасибо за внимание!