УДК 004.912: 81-114.2 https://doi.org/10.33619/2414-2948/40/27

ТЕСТИРОВАНИЕ ПРОГРАММЫ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА ЕСТЕСТВЕННОГО ЯЗЫКА

©Сатыбаев А. Д., *д-р физ.-мат. наук, Ошский технологический университет им. М. М. Адышева, г. Ош, Кыргызстан, abdu—satybaev@mail.ru*
©Кочконбаева Б. О., *Ошский технологический университет им. М. М. Адышева, г. Ош, Кыргызстан, buajar@mail.ru*

TESTING OF THE PROGRAM OF THE MORPHOLOGICAL ANALYZER OF A NATURAL LANGUAGE

©Satybaev A., *Dr. habil, Osh Technological University named by M.M. Adyshev, Osh, Kyrgyzstan, abdu-satybaev@mail.ru*
©Kochkonbaeva B., *Osh Technological University named by M.M. Adyshev, Osh, Kyrgyzstan, buajar@mail.ru*

Аннотация. Данная статья посвящена описанию разработки интегрированных лингвистических моделей данных и программных модулей для морфологического анализа кыргызских словоформ. Созданная модель и базы данных могут быть реализованы не только в технологии семантического поиска для повышения функциональности поисковых систем, но и в других системах обработки тюркских языков.

Abstract. This article describes the development of integrated linguistic data models and software modules for the morphological analysis of Kyrgyz word forms. The created model and databases can be implemented not only in semantic search technology to enhance the functionality of search engines, but also in other systems for processing Turkic languages.

Ключевые слова: функция, система, базы данных, интерфейс, морфология, кыргызский язык, алгоритм, словоформа.

Keywords: function, system, data base, interface, morphology, Kyrgyz language, algorithm, word form.

Введение

На современном этапе развития науки и техники предпочтительны процессы обработки информации, которые занимают лидирующие позиции в производстве и охватывают все сферы человеческой деятельности. Методы и средства обработки информации на естественном языке становятся все более и более важными — от простейших систем подготовки документов до информационно-поисковых систем, систем машинного перевода и программ общения на естественном языке. Чрезвычайно широкий спектр приложений, в связи с обработкой текстов на естественном языке. Глубина их проникновения в структуру текста также различна.

Морфологический анализ — это процесс сегментации слов в морфемы или анализ процесса формирования слов. Это основной шаг для различных типов текстового анализа любого языка. Морфологический анализатор берет слово в качестве входных данных и вырабатывает корень, а его грамматические функции — как результат. Например, на основе

морфологичекого анализа в среде RAD Studio XE3 была составлена система анализатора NLP.

Структура системы

Система состоит из базы данных и интерфейса пользователя, модуля морфологического анализа и статистического анализа. Концептуальная схема программы приведен на Рисунке 1.

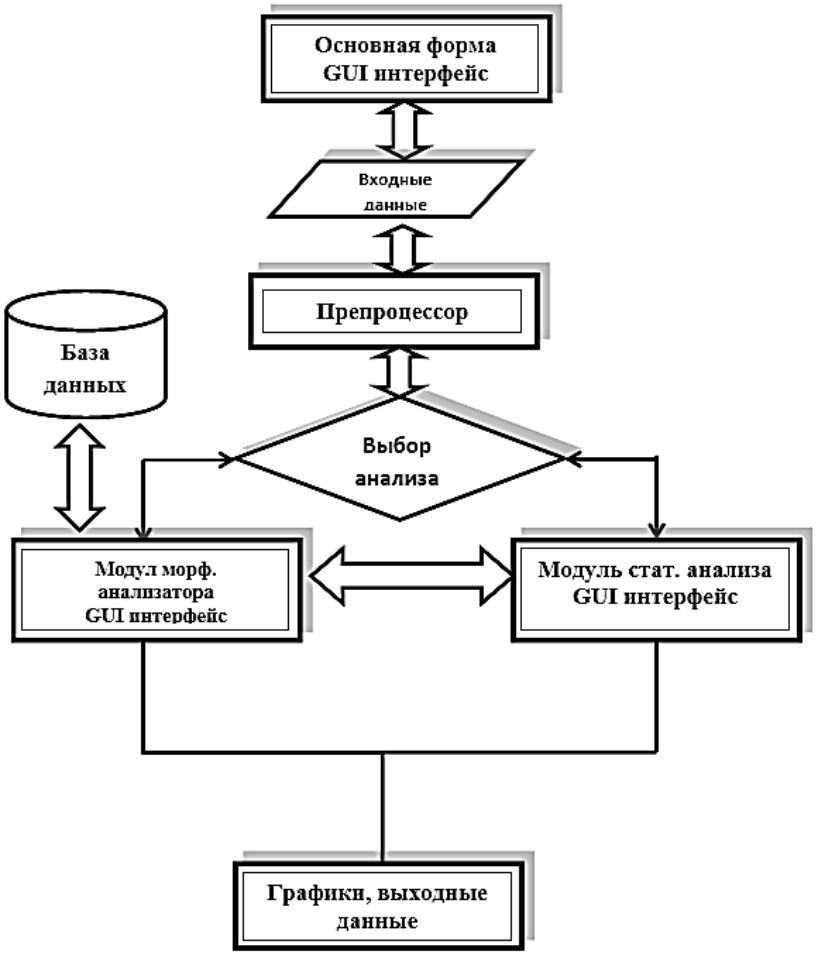

Рисунок 1. Концептуальная схема системы.

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.

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 (Рисунок 2), который позволяет производить запросы к базе данных.

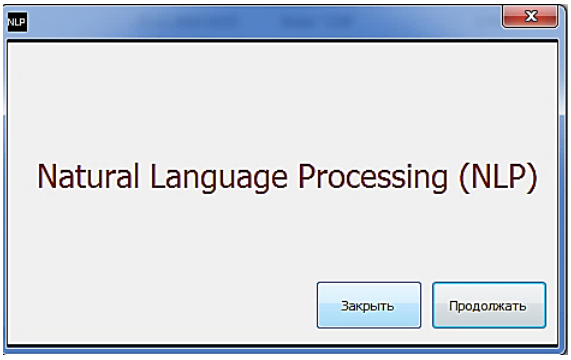

Рисунок 2. Интерфейс системы.

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 представляют собой множества аффиксов, образованные по морфонологическим правилам кыргызского языка (Рисунок 3).

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)

Рисунок 3. а) база данных аффиксов; б) база данных основ.

Теоретически эти цепочки, образуемые словоизменительными аффиксами, в агглютинативных тюркских языках могут иметь бесконечную длину. Однако при создании базы данных было принято 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 (Рисунок 4).

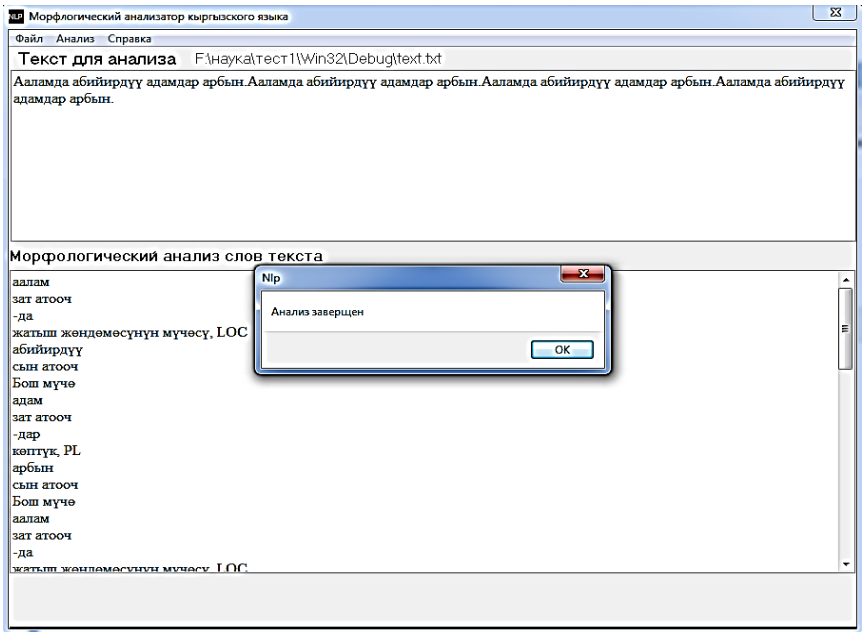

Рисунок 4. Результат работы программы анализатора.

А также система имеет модуль статистического анализа, где вычисляется частота выполнения проверки и разбиения каждого слова (Рисунок 5).

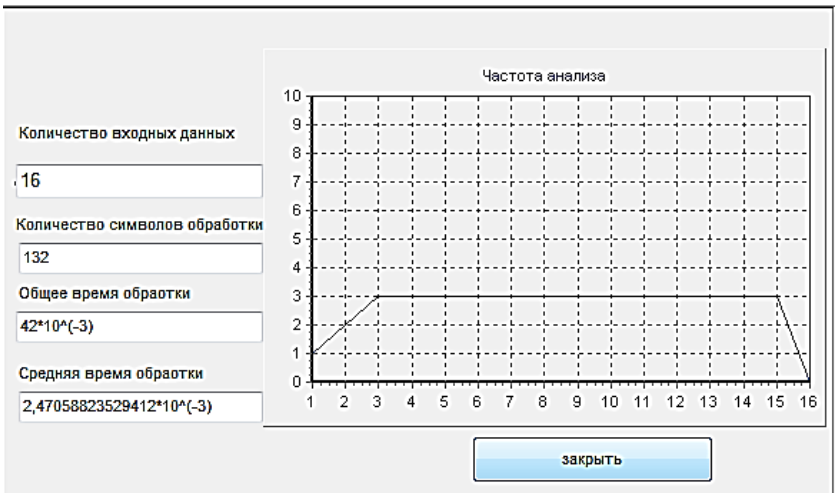

Рисунок 5. Результат статистического анализа.

Заключение

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.

Список литературы:

1. Мансуров К. Т., Кочконбаева Б. О., Ергешова Г. Программирование в среде Delphi 2006. Ош, 2009.
2. Gatiatullin A., Ayupov M. Modifications of morphological analysis programs for the problems of multilingual search // Proceedings of the International Conference “Turkic Languages Processing” Kazan: TurkLang, 2015. С. 120-126.
3. Kochkonbaeva B., Aldosova A. Automatic processing of text in natural language // Бюллетень науки и практики. 2018. Т. 4. №7. С. 216-221.

References:

1. Mansurov, K. T., Kochkonbaeva, B. O., Ergeshova, G. Programirovanie v srede Delphi 2006. Osh, 2009. (in Russian).
2. Gatiatullin, A., & Ayupov, M. (2015). Modifications of morphological analysis programs for the problems of multilingual search. *In: Proceedings of the International Conference “Turkic Languages Processing”. Kazan, TurkLang, 2015,* 120-126.
3. Kochkonbaeva, B., & Aldosova, A. (2018). Automatic processing of text in natural language. *Bulletin of Science and Practice, 4*(7), 216-221.

Работа поступила
в редакцию 11.02.2019 г.

Принята к публикации
16.02.2019 г.

Ссылка для цитирования:
Сатыбаев А. Д., Кочконбаева Б. О. Тестирование программы морфологического анализатора естественного языка // Бюллетень науки и практики. 2019. Т. 5. №3. С. 215-219. https://doi.org/10.33619/2414-2948/40/27.

Cite as (APA):
Satybaev, A., & Kochkonbaeva, B. (2019). Testing of the program of the morphological analyzer of a natural language. *Bulletin of Science and Practice, 5*(3), 215-219. https://doi.org/10.33619/2414-2948/40/27. (in Russian).