РАЗРАБОТКА ЛИНГВИСТИЧЕСКОГО ПРОЦЕССОРА ТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ

К. Ч. Койбагаров, Р. Р. Мусабаев, М. Н. Калимолдаев

Институт информационных и вычислительных технологий Комитета науки Министерства образования и науки Республики Казахстан,
050010, Алма-Ата, Республика Казахстан

УДК 681.5

Настоящая работа посвящена описанию модуля лексико-морфологического анализа слов казахского языка, который будет использоваться в качестве инструмента облачного веб-сервиса. В работе обоснованы способы представления морфологической информации и хранения и методы доступа к словам словаря. Описан принцип работы морфологического анализатора. Дан обзор методов анализа на основе теории конечных автоматов, показаны особенности и характеристики данного подхода в представленном анализаторе.

Ключевые слова: морфологический анализ, детерминированный конечный автомат, закон сингармонизма.

In this paper, we introduce the morphological parser module words of the Kazakh language. We present finite-state implementation of a morphological parser in agglutinative languages such as the Kazakh language. The main contribution of this paper is to give a thorough description of a perspective for stemming which can also be generalized to apply to other agglutinative languages like Finnish, Hungarian, Estonian, Czech and Turkic.

Key words: linguistic processor, token, finite-state machine, morphological parser, lexical analysis, suffixes, state diagram, law synharmonism.

Введение. Лавинообразный рост ИТ-технологий и количества информации, которую приходится анализировать человеку, сделали насущной проблему организации управления данными одной из наиболее важных для многих областей жизнедеятельности. Без современных средств анализа информации невозможно принятие своевременных и обеспечивающих конкурентоспособность решений вплоть до уровня крупных организаций и государств.

Популярным способом представления информации являются неструктурированные текстовые документы (блоги, форумы, интернет-сайты, СМИ и т. д.). Информация в таких документах представлена в слабоструктурированном виде, что существенно усложняет ее обработку. Огромные массивы накопленных текстовых данных должны помогать принимать правильные решения, анализировать обстановку в обществе, ситуацию на рынке и многие другие параметры. Однако для анализа и принятия решений многим частным компаниям просто не хватает инструментов. Человеческий язык отличается множеством оборотов речи и прочих эмоциональных проявлений, непонятных машинной логике. Применение ИТ-инструментов становится актуальным в связи с растущим объемом текстового контента, является особенно интересным для решения актуальных задач, таких как выявление мнений, анализ отзывов, оценка общественных настроений. Для автоматического

анализа таких данных в Институте информационных и вычислительных технологий МОН РК ведется разработка лингвистического процессора, на базе которого будет создан веб-сервис. Таким образом, разработка инструментов текстового анализа и представление их в виде облачных сервисов позволит создавать интеллектуальные веб-приложения.

В настоящее время разработано большое количество библиотек для анализа текстов на естественном языке, содержащих наборы базовых алгоритмов для анализа текстов, в основном на английском языке. Наиболее известными из них являются OpenNLP, NLTK [1], LingPipei. Для русского языка наиболее известным пакетом инструментов является АОТ [2], а также решения компаний ABBYY, RCO, IBM и др. Для казахского языка аналогичных пакетов программ по обработке текстов не обнаружено.

Анализ исследования открытых публикаций в области технологий лингвистического анализа словоформ казахского языка показывает, что исследований в данной области мало.

В 1970–2000 годах публикации в области морфологии казахского языка носили в основном теоретический характер. С 2006 года появились публикации в зарубежных журналах: Jonathan North Washington „A Novel Approach to Delineating Kazakh's Five Present Tenses: Lexical Aspect“ (2006); A. Dzhubanov, B. Khasanov „Computational description of the kazakh language“ (2007); B. J. Bayachorova, P. Pankov „Independent Computer presentation of a natural language“ (2009); G. Altenbek and Wang Xiao-long „Kazakh Segmentation System of Inflectional Affixes“ (2010); а также работа Sharipbaev A. A., Bekmanova G. T., Ergesh B. J., Buribaeva A. K., Karabalaeva M. H. „The intellectual morphological analyzer based on semantic network“ (2012).

Все перечисленные работы затрагивают некоторые аспекты в области морфологии и синтаксиса казахского языка и носят теоретический характер исследований. Как ни странно, для казахского языка разработанной технологии морфоанализа реализовано не было, по крайней мере, в открытых источниках информации о ней найти не удалось. Так как казахский язык относится к тюркской группе языков, можно отметить похожие работы для турецкого и башкирского языков: Орехов Б. В., Слободян Е. А. „Проблемы автоматической морфологии агглютинативных языков и парсер башкирского языка“; B. Taner Dincer, Bahar Karaoglan „Stemming in Agglutinative Languages: A Probablistic Stemmer for Turkish“ [Taner Dincer].

Отсутствие технологии анализа текстов казахского языка препятствует дальнейшим исследованиям в области синтаксиса и семантики. Таким образом, актуальным является создание лексико-морфологического анализа слов казахского языка с высокой скоростью обработки слов, применимого для различных систем по обработке текстов на казахском языке.

Особенности казахской морфологии. Казахский язык принадлежит к тюркской семье языков, куда относятся также узбекский, кыргызский, татарский, башкирский, азербайджанский, турецкий и др. Казахский язык относится к классу агглютинативных языков. Для агглютинативных языков характерно последовательное присоединение различных формообразующих суффиксов или окончаний, несущих грамматическое значение, к неизменяемому корню или основе, являющихся носителями лексического значения.

Порядок добавления аффиксов строго определен (рис. 1). Например, для имен существительных к основе слова вначале добавляется суффикс, окончание множественного числа, затем притяжательное окончание, далее следует падежное окончание и последним окончание формы спряжения (добавляется только к одушевленным существительным) [3].

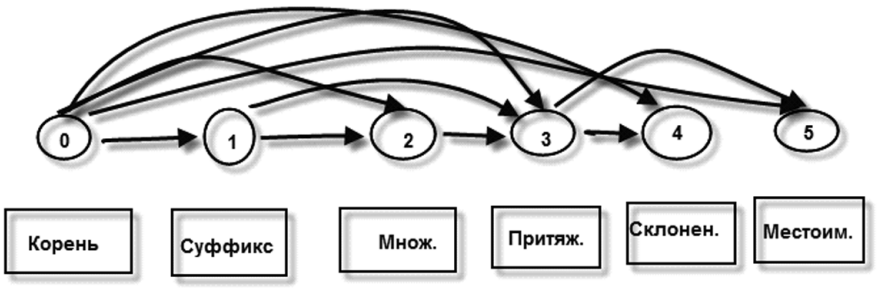

Рис. 1. Правило присоединения аффиксов для имен существительных

В казахском языке от одного корня слова можно сгенерировать большое количество словоформ, в особенности для глаголов. Например:

Қанағаттандырылмағандықтарыңызбенен.

Мы можем выделить аффиксы из данной словоформы:

қанағат+тан+дыр+ыл+ма+ған+дық+тар+ыңыз+бенен.

Кроме определенных правил присоединения аффиксов для каждой части речи, нужно также учитывать и фонетические особенности.

Специфика казахского языка. Для казахского языка существуют законы сингармонизма для гласных и согласных — закон созвучия (гармонии) звуков основы слова и аффиксов. Гармонируют гласные по принципу твердости-мягкости и согласные — конечный звук корня и первый звук аффикса.

Учитывая вышесказанное, мы определили три закона сингармонизма, которые нужно учитывать при морфологическом разборе слова.

1) Сингармонизм гласных. В казахском языке гласные относятся к двум группам: мягкие гласные и твердые гласные.

2) Прогрессивная ассимиляция согласных. Если слова оканчиваются на глухие согласные, то прибавляемые к слову аффиксы начинаются с глухих согласных. Если последний слог слова звонкий, сонорный или гласный, то прибавляемые окончания начинаются со звонкого согласного.

3) Регрессивная ассимиляция. Если последний слог оканчивается на глухую согласную „к“, „қ“, „п“, а к нему добавляется окончание, которое начинается с гласной, то глухие согласные становятся звонкими: „к – г“, „қ – ғ“, „п – б“.

Помимо трех основных правил сингармонизма, необходимо учитывать следующие правила исключения.

1) Правило удаления глухой согласной в прибавляемом аффиксе, если в окончании слова присутствуют две глухие согласные. Например: журналист + тер → журналистер, экстремист+тер → экстремистер.

2) Закон сингармонизма не соблюдается для аффиксов:

Мен, пен, бен — қал**а**мм**е**н;

Нікі, дікі, тікі — бал**а**н**і**кі;

для заимственных слов с окончаниями:

Рк, нк, кс, кт — п**у**кнт**е.**

1) Правило выпадения гласной „і“, „ы“ в корне слова при добавлении притяжательного аффикса „і“, „ы“. Например: Әріп — әрпі, қауіп — қаупі, қойын — қойны, Ерін — ерні, құлық — құлкы.

2) **Морфологический анализ.** Среди методов морфологического анализа, использующихся в лингвистических процессорах, можно выделить методы с декларативной и с процедурной ориентацией.

Декларативный подход. Для методов декларативной ориентации характерно наличие полного словаря всех возможных сочетаний аффиксов. При этом каждый аффикс снабжается полной и однозначной морфологической информацией, куда входят переменные морфологические параметры. Задача морфологического анализа в этом случае сводится к поиску нужной словоформы в словаре основ слов и аффикса в словаре аффиксов, копированию морфологической информации, соответствующей найденной словоформе, в программу. Так как количество различных словоформ каждого слова довольно велико, декларативный метод требует значительных затрат памяти вычислительной системы, что сопровождается рядом трудностей, связанных с созданием и поддержкой словаря, а также с избыточностью информации. К достоинствам данного метода следует отнести высокую скорость анализа и универсальность по отношению к множеству всех возможных словоформ.

Процедурный подход. В процедурных методах каждое слово разделяется на основу слова и цепочку основ аффиксов. Для этого используются словари основ слов и основ аффиксов. Основной критерий при разбиении слова на основу и аффикс: основа должна оставаться неизменной во всех возможных словоформах данного слова, кроме глаголов, у которых нужно отбросить последнюю букву „у“, „ю“. Поскольку большое количество слов казахского языка использует одни и те же аффиксы, то суммарный объем словаря основ и словаря аффиксов оказывается значительно меньше, чем объем полного словаря всех сочетаний аффиксов, используемого в декларативных методах. Однако процедура морфологического анализа усложняется: теперь из словаря основ необходимо поочередно выбирать все основы, совпадающие с начальными буквами анализируемого слова и для каждой такой основы перебирать все возможные для нее аффиксы. В случае точного совпадения очередного варианта „основа+аффикс1+аффикс2...“ с анализируемым словом вариант анализа считается успешным, и в программу передается морфологическая информация, соответствующая данной основе и данному аффиксу. При этом, как правило, постоянные морфологические параметры определяются основой слова, а переменные — аффиксом.

Для процедурных методов время анализа одного слова может быть существенно выше, но объем используемых словарей в небольших системах позволяет загружать словари целиком в оперативную память. Морфологический анализатор содержит также модули по выделению временных групп в тексте (времени и даты), организации, персоналий и географических имен.

Гибридный подход. Из-за особенности морфологии агглютинативных языков нами было предложено использовать преимущество двух подходов: декларативного и процедурного метода. Для процессора важнейшими свойствами являются скорость анализа входной последовательности символов и компактность. Эти характеристики особенно важны для анализатора при обработке больших объемов текста.

На первом этапе используется декларативный метод анализа словоформ, и в случае положительного результата анализ данной словоформы прекращается. В противном случае

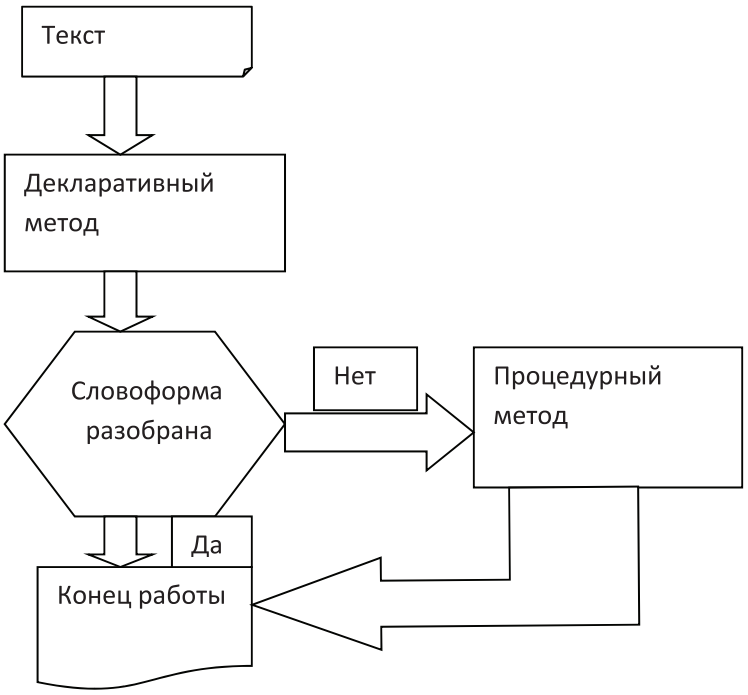

Рис. 2. Схема работы морфологического анализатора

словоформа передается для дальнейшего разбора в модуль, использующий процедурный подход (рис. 2).

Способ реализации модуля с декларативным подходом. Постоянную и переменную морфологическую информацию будем хранить в двух словарях. Постоянная морфологическая информация будет храниться в словаре лемм, а переменная часть — в словаре окончаний. Для решения проблемы законов сингармонизма была выбрана модель хранения возможного окончания с двумя предшествующими буквами неизменяемой части слова. Так, например, словоформа „кітап**тар**“ порождает правило, разрешающее отщепление окончания **-тар** при условии, что ему предшествует последовательность **-ап**.

Для получения словаря окончаний (аффиксальных соединений), нами был разработан программный модуль обработки массивов полнотекстовой информации, который, разбив текст на слова, выполнял обработку очередной словоформы точным морфологическим анализатором [4], имея базу более 150 000 основ слов и 550 основ аффиксов и распознавая более 20 млн различных словоформ казахских слов. В результате работы программы из словоформ выделялась их основа, то есть часть слова, остающаяся неизменной при склонении; выделенное таким способом окончание вместе с последними двумя символами формальной основы поступало в таблицу окончаний.

Для описания словоизменения слово разбивается на две таблицы:

– основа — неизменяемая графическая часть слова;

– окончание — аффиксальное соединение.

Для каждого слова после выделения всех изменяемых частей в словарь заносится часть речи, ссылка на таблицу окончаний. В таблице окончаний также хранятся грамматические характеристики слов. По полученной таблице слов и окончаний строится конечный детерминированный автомат с диаграммой переходов в виде дерева.

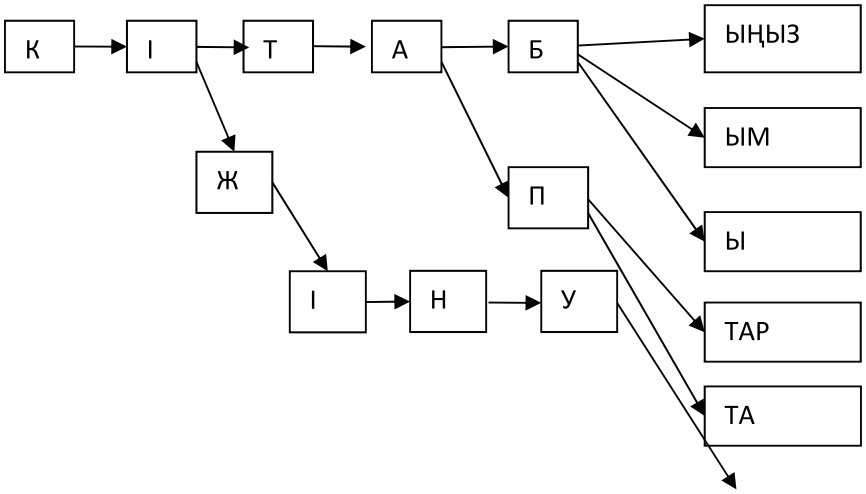

Рис. 3. Диаграмма переходов конечного автомата

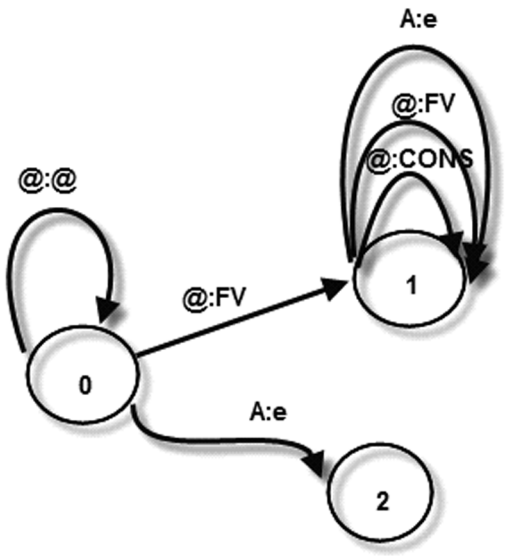

Рис. 4. Автомат, реализующий фонетические правила

Приведем пример для слов „кітабыңыз“, „кітабым“, „кітабы“, „кітаптар“, „кітапта“ (рис. 3).

Для получения словаря окончаний мы с помощью разработанного нами точного морфологического анализатора протестировали корпус текстов, содержащий приблизительно 500 000 слов. В результате работы анализатора получили около 6500 окончаний.

Способ реализации модуля с процедурным подходом. Для моделирования правил соединения аффиксов (рис. 1) мы используем детерминированный конечный автомат (ДКА) [5]. Для разработки морфологического анализатора необходимы словарь основ слов с грамматическими характеристиками, словарь основ аффиксов с грамматическими характеристиками, правила соединения аффиксов для каждой части речи, а также фонети-

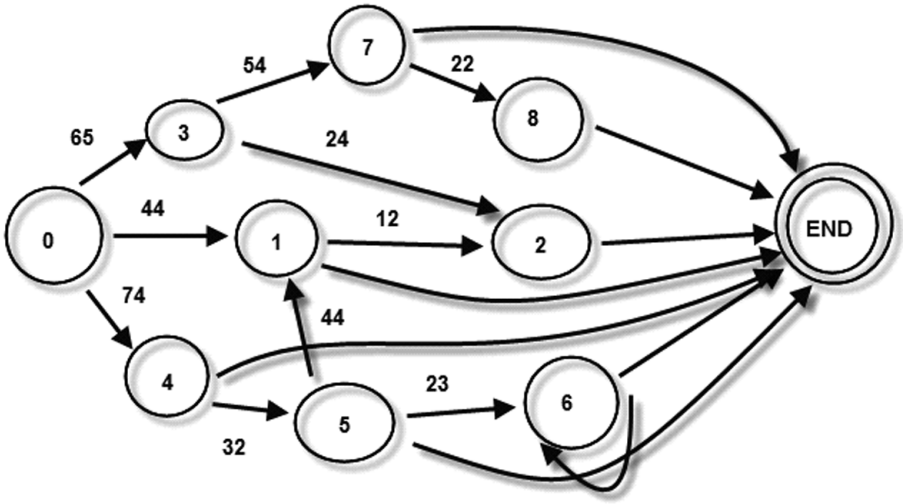

Рис. 5. Диаграмма состояний детерминированного автомата, распознающего существительное

ческие правила (законы сингармонизма). Правила соединения аффиксов и фонетические правила имплементированы в детерминированный конечный автомат.

Для использования фонетических правил в работе [6] предлагается использовать двухуровневый морфологический формализм. Фонетические правила представляются в виде конечного автомата (рис. 4).

A:e -> @:FV [@:CONS |@:“]

Символ „A“ представляет гласные согласные в аффиксе, „FV“ — мягкие гласные (ә, ө, ү, і, е), „CONS“ представляет собой множество согласных, „@“ представляет собой любые символы алфавита.

Учитывая вышесказанное, мы реализовали двухуровневый морфологический анализатор в виде двух модулей. На первом уровне окончание разбирается с помощью автомата, реализующего правила соединения аффиксов, а на втором уровне автомат проверки фонетических правил.

Для формализации правил соединения аффиксов друг к другу мы собрали таблицу из 550 аффиксов.

Полученные аффиксы сгруппировали в классы по признакам. Всего получилось 78 классов. Далее составили правила соединения аффиксов для семи частей речи казахского языка (существительное, глагол, прилагательное, наречие, местоимение, модальный глагол, числительное). Эти правила представлены в виде детерминированных конечных автоматов. Например, модель автомата, распознающего существительное, представлена на рис. 5.

Цифры в кружочках представляют собой состояния, а цифры на дугах — это номера классов, содержащих определенный набор аффиксов, которые переводят автомат из одного состояния в другое. Таким же образом мы сделали автоматы для семи частей речи.

Алгоритм морфологического анализа. Нами был предложен следующий алгоритм морфологического анализатора казахских словоформ для процедурного метода.

Таблица

Таблица аффиксов

Аффикс	Описание	Часть речи	Номер
ын	Возвратный залог	Глаг.Зал.	21
ін	Возвратный залог	Глаг.Зал.	21
н	Возвратный залог	Глаг.Зал.	21
дыр	Понудительный залог	Глаг.Зал.	22
дір	Понудительный залог	Глаг.Зал.	22
тыр	Понудительный залог	Глаг.Зал.	22
тір	Понудительный залог	Глаг.Зал.	22
т	Понудительный залог	Глаг.Зал.	22
ғыз	Понудительный залог	Глаг.Зал.	22
ғіз	Понудительный залог	Глаг.Зал.	22
қыз	Понудительный залог	Глаг.Зал.	22
кіз	Понудительный залог	Глаг.Зал.	22
ыр	Понудительный залог	Глаг.Зал.	22
ір	Понудительный залог	Глаг.Зал.	22
...	...	...	...

Шаг 1. Текстовый файл подается на вход лексического анализатора. Текст разбивается на предложения. Из предложения выделяются слова, которые подаются на вход морфологического анализатора.

Шаг 2. Из цепочки слов выбирается в цикле слово.

Шаг 3. Проверяется слово в словаре основ слов. Если слово распознано, то переходим к шагу 2.

Шаг 4. Слово побуквенно считывается справа налево. Полученное слово ищется в словаре основ слов. Если слово найдено, то остаток слова считается аффиксальным соединением. В найденном слове считываем часть речи.

Шаг 5. Выбираем распознающий автомат, соответствующий части речи обнаруженного слова в словаре. Если автомат не распознал окончание, то продолжаем разбор слова — шаг 4.

Шаг 6. Проверка окончания на соответствии законам сингармонизма. Если проверку не прошел, продолжаем разбор слова — шаг 4. Если проверку прошел, то записываем грамматическую характеристику словоформы.

Шаг 7. Если все слова прошли разбор, то выход их программы. Если нет, то шаг 2.

Заключение. Алгоритм работы морфологического анализатора основан на реализации процедуры сегментации основы слова, суффикса и аффиксов из лексемы и определении его морфологической информации.

Нами был собран словарь основ слов объемом 85 000, словарь фамилий и имен (60 000), словарь географических названий (5 000), а также 550 основ аффиксов и 6500 окончаний составляют основу работы морфологического анализатора.

С помощью разработанного нами веб-паука из казахстанских сайтов скачано около 2 Гб текстов на казахском языке, примерно 200 млн токенов и около 70 млн слов.

В настоящее время лингвистический процессор поддерживает следующие стандартные методы:

1) определение границ предложений в тексте;
2) определение границ отдельных слов, или токенов, в предложении;
3) определение частей речи слов;
4) приведение слов к нормальной форме.

Разработка лингвистического процессора ведется на языке C#, и результаты наших исследований можно увидеть на сайте института.

Список литературы

1. Steven Bird, Ewan Klein, Edward Loper, and Jason Baldridge. Multidisciplinary instruction with the Natural Language Toolkit / InProceedings of the 3-d Workshop on Is. in Teach. Computat. Ling. (TeachCL '08). Association for Computational Linguistics, Stroudsburg, PA, USA, 2008. P. 62–70.

2. Игорь Ножов. Морфологическая и синтаксическая обработка текста (модели и программы), тезисы диссертации. 2003.

3. Бекманова Г. Т. Некоторые подходы к проблемам автоматического словоизменения и морфологического анализа в казахском языке // Вестник Восточно-Казахстанского государственного технического университета им. Д. Серикбаева. Усть-Каменогорск. 2009. № 4. С. 192–197.

4. Койбагаров К. Ч., Мусабаев Р. Р., Кулманов С. К. Разработка алгоритмов автоматического анализа словоформ казахского языка // Труды Междунар. научно-теоретич. конф. „Современное казахское языкознание“. Алма-Ата. 2012. С. 83.

5. Кудрявцев В. В., Алешин С. В., Подколзин А. С. Введение в теорию автоматов. М.: Наука, 1985.

6. Koskenniemi, K. A.General Computational Model for Word-form Recognition and Production. / 22 An. Meeting on Association for Computational Linguistics. 1984. P. 178–181.

Койбагаров Кайрат Чапденович — научный сотрудник Института проблем информатики и управления МОН РК;
Мусабаев Рустам Рафикович — канд. техн. наук, зав. лабораторией Института проблем информатики и управления МОН РК;
Калимолдаев Максат Нурадилович — д-р физ.-мат. наук, проф., ген. дир., зав. лаб. математического моделирования и кибернетики Института проблем информатики и управления МОН РК, тел.: +7 (727) 272-37-11, e-mail: mnk@ipic.kz

Дата поступления — 10.09.2014