УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2020 Том 20 № 4 ISSN 2226-1494 http://ntv.itmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July–August 2020 Vol. 20 No 4 ISSN 2226-1494 http://ntv.itmo.ru/en/
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

УДК 004.912 doi: 10.17586/2226-1494-2020-20-4-545-551

ЗАДАЧА НОРМАЛИЗАЦИИ СЛОВ КАЗАХСКОГО ЯЗЫКА

Д.Р. Рахимова[a,b], А.О. Турганбаева[a]

[a] Казахский Национальный Университет имени Аль Фараби, Алматы, 050040, Казахстан
[b] Институт информационных и вычислительных технологий Алматы, 050000, Казахстан
Адрес для переписки: diana.rakhimova@kaznu.kz

Информация о статье
Поступила в редакцию 01.06.20, принята к печати 25.06.20
Язык статьи — русский

Ссылка для цитирования: Рахимова Д.Р., Турганбаева А.О. Задача нормализации слов казахского языка // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20. № 4. С. 545–551. doi: 10.17586/2226-1494-2020-20-4-545-551

Аннотация

Предмет исследования. Рассмотрены модели и существующие алгоритмы нормализации слов естественных языков. Описаны алгоритмы автоматического выделения основ для ряда естественных языков и возможные пути синтеза нормальной формы слова для казахского языка. **Цель.** Создание полной классификации системы окончаний для казахского языка. Разработка алгоритма нормализации слов на основе предложенного подхода классификации окончаний и суффиксов. **Методология.** Проведен анализ словообразования с помощью окончаний для всех частей речи казахского языка, на основе выполненной работы представлена классификация окончаний и суффиксов. Рассмотрены возможные варианты размещений типов окончаний и суффиксов. Общее количество возможных суффиксов составляет 26 526 единиц, окончаний – 3 565 единиц. Все приведенные типы являются лексически и семантически допустимыми, но некоторые из них не применяются. В базу аффиксов добавлены только те, которые наиболее часто применяются. С помощью множеств представлено, в каком порядке к основе добавляется аффиксы. Это нужно для того, чтобы правильно выделить основу. В работе не рассматриваются словообразующие суффиксы, так как они меняют основу слова и контекст значения. В основном к существительным добавляются словообразующие суффиксы. **Основные результаты.** Разработана полная система классификации окончаний и суффиксов казахского языка. Построены детерминированные конечные автоматы для различных частей речи с использованием всевозможных вариантов добавления суффиксов и окончаний с учетом морфологических и лексических свойств грамматики казахского языка. Разработан алгоритм стеминга с использованием разработанной системы классификации окончаний казахского языка. Реализована система нормализации, доказывающая работоспособность разработанного алгоритма без словаря. Алгоритм протестирован на корпусе казахского языка. В заданном корпусе изначально были убраны знаки пунктуации и стоп-слова. **Практическая значимость.** Результаты работы могут найти применение при анализе текста, нормализации (лемматизации) текста, а также в информационно-поисковых системах, в машинном переводе казахского языка и других прикладных задачах.

Ключевые слова
обработка естественного языка, казахский язык, система окончаний, нормализация, алгоритм стеминга

Благодарности
Исследование выполнено при поддержке Министерства образования и науки Республики Казахстан в рамках научного проекта AP 05132950.

doi: 10.17586/2226-1494-2020-20-4-545-551

NORMALIZATION OF KAZAKH LANGUAGE WORDS

D.R. Rakhimova[a,b], A.O. Turganbaeva[a]

[a] Al-Farabi Kazakh National University, Almaty, 050040, Republic of Kazakhstan
[b] Institute of Information and Computing Technologies, Almaty, 050000, Republic of Kazakhstan
Corresponding author: diana.rakhimova@kaznu.kz

Article info
Received 01.06.20, accepted 25.06.20
Article in Russian

For citation: Rakhimova D.R., Turganbaeva A.O. Normalization of Kazakh language words. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2020, vol. 20, no. 4, pp. 545–551 (in Russian). doi: 10.17586/2226-1494-2020-20-4-545-551

Abstract

Subject of Research. Models and existing algorithms for normalization of natural language words are considered. The paper describes algorithms for automatic selection of the basic principles for a number of natural languages and possible ways of the normal word form synthesis for the Kazakh language. The research is aimed at creation of a complete classification for the Kazakh language ending system and development of a normalization algorithm for words based on the proposed classification approach for endings and suffixes. **Method.** Word formation analysis by applying endings for all Kazakh language parts of speech was carried out; a classification of endings and suffixes was presented. The paper discusses all kinds of placement options for endings and suffixes. The total number of various suffixes is 26 526 units and the endings is 3 565 units. All considered types are lexically and semantically valid, but some of them are not applicable. Only those, that are most commonly used, are added to the affix base. The order, that the affixes are added to the base, is presented using sets. Thus, the base is correctly selected. The study does not examine word-forming suffixes, as they change the word stem and contextual interpretation. Basically, word-forming suffixes are added to nouns. **Main Results.** A complete classification system for endings and suffixes of the Kazakh language has been developed. Deterministic finite automata for various parts of speech are created using all possible options, adding suffixes and endings, taking into account the morphological and lexical features of the Kazakh language grammar. A lexicon-free stemming algorithm is developed using the proposed classification system for endings of the Kazakh language. A normalization system has been implemented, proving the operability of the developed algorithm without a dictionary. The algorithm implementation was tested on the Kazakh language corpus. Punctuation and stop words were initially removed from the specified corpus. **Practical Relevance.** The results of the work can find application in the text analysis and normalization (lemmatization), as well as in information retrieval systems, in machine translation from the Kazakh language, and other applied problems.

Keywords

natural language processing, Kazakh, ending system, normalization, stemming algorithm

Acknowledgements

The study was supported by the Ministry of Education and Science of the Republic of Kazakhstan within the framework of the AP05132950 scientific project.

Введение

В настоящее время активно создаются различные интеллектуальные и мобильные системы, связанные с обработкой естественного языка (ОЕЯ). К сожалению, вопросы текстовой обработки казахского языка слабо развиты, что препятствует развитию информационных технологий, и связано:

1) со спецификой казахского языка как языка со сложной морфологией;
2) с отсутствием электронных ресурсов для изучения казахского языка в этой области.

Тем не менее вопросы обработки текстов на казахском языке на практике являются очень актуальными. Важной проблемой является проблема быстрого поиска конкретных слов в документах. Один из способов быстрого поиска слов, заключается в поиске основы слова среди ключевых слов документов, позволяющей выбрать соответствующий документ как желаемый. Одним из важных процессов в прикладных системах ОЕЯ, таких как информационный поиск, машинный перевод и др., является нормализация (лемматизация), т. е. приведение слова к изначальной основе.

Различными учеными и научными группами выполнен анализ, и рассмотрены разные подходы по нормализации казахского языка. По направлению сегментации аффиксов казахского языка можно рассматривать работу [1], где проанализирована морфемная структура в корпусе казахского языка, и изучено извлечение основ и сегментации аффикса. Сначала устанавливается finite-state machine (FSM — конечный автомат) флективных аффиксов, а затем проводится сегментация флективных аффиксов.

В работе автором в качестве флективных аффиксов указаны четыре вида окончаний: множественного числа, притяжательные, падежные, личные. По направлению морфологического анализа казахского языка можно отнести работы [2, 3]. В работе [2] построен морфологический анализатор с использованием двухуровневого морфологического подхода с инструментами конечного состояния Xerox, и представлена реализация морфологического анализатора на основе правил. В работе [3] представлены морфологические особенности казахского и турецкого языков. Проведено сравнение онтологии, разработана единая система символов морфологических признаков, морфологические правила казахского и турецкого языков записывались через новую систему символов. Унифицированный морфологический анализатор разработан на основе алгоритма общего морфологического анализа. В работе [4] представлен подход по нормализации основных типов окончания казахского языка. В вышеупомянутых работах в основном исследовался определенный (ограниченный) класс окончаний казахского языка, но не были представлены сложные формы изменения языка, с помощью анализа и генерации окончаний (суффиксов и аффиксов) казахского языка, что не покрывает ее полноту. В настоящей работе авторы представляют новый подход по классификации окончаний казахского языка, полноту покрытия и практическое применение.

Классификация окончаний казахского языка

В казахском языке в состав аффиксов входят суффиксы и окончания. Авторами под руководством профессора У.А. Тукеева [5] разработана полная система окончаний казахского языка. Типы окончаний разделены на два класса:

1) система окончаний к именным основам слов;
2) глагольным основам казахского языка.

Приведены четыре типа системы окончаний к именным основам слов казахского языка:
1) окончания множественного числа (*K*);
2) притяжательные окончания (*T*);
3) падежные окончания (*C*);
4) личные окончания (*J*).

Основа (stem) (*S*). В работе рассмотрены возможные варианты размещений типов окончаний, которые определены формулой: $A_n^k = n!/(n–k)!$

Имеется *n* различных объектов. Будем выбирать из них *k* объектов и переставлять всеми возможными способами между собой (т. е. меняется и состав выбранных окончаний, и их порядок). Получившиеся комбинации называются размещениями из *n* объектов по *k*. В результате расчета количество возможных размещений равно 64, но некоторые из них являются семантически не допустимыми. Семантически допустимыми будут следующие: *KT*, *TC*, *CJ*, *KC*, *TJ*, *KJ*, *KTC*, *KTJ*, *TCJ*, *KCJ*, *KTCJ*. Итого, допустимых размещений из одного типа — 4, из двух типов — 6, из трех типов — 4, из четырех типов — 1. Суммарное число типов допустимых размещений в словах с именными основами — 15. По полученным типам окончаний определены формы окончаний и их количество. Так, для частей речи с именными основами количество окончаний равно 1 213 (учтены все варианты множественного числа), а количество окончаний частей речи с глагольными основами составляет: глаголы – 432, причастия – 1 582, деепричастия – 48, наклонения – 240, залоги – 80 (они не являются окончаниями, а являются суффиксами, образующими формы глагола, поэтому далее их будем рассматривать как суффиксы). Итого всего 3 565 единиц окончаний.

Классификация суффиксов казахского языка

В казахском языке существуют два вида суффиксов:
1) словообразующие (*TdJr*);
2) формообразующие (*TrJr*).

Словообразующие суффиксы служат для образования новых слов (меняется смысл слова). Например: *ән* — песня, *ән-ші* — певец; *ақыл* — ум, *ақыл-ды* — умный.

Формообразующие суффиксы служат для образования форм слова. Например: *жеті* — семь, *жеті-нші* — седьмой; *қатты* — твердый, *қатты-рақ* — тверже; *оқы* — читай, *оқы-лды* — прочитано.

В свою очередь, формообразующие суффиксы подразделяются на модифицированные и грамматические суффиксы.

К модифицированным суффиксам относятся: производные суффиксы существительного; суффиксы степени сравнения имен прилагательных; суффиксы порядковых имен числительных; суффиксы глаголов, формирующих залог, усилительный и отрицательный виды глагола.

Суффиксы отрицательного глагола и производные суффиксы существительного не будут добавлены в базу аффиксов. Например, в слове бармаңыздар (бар + ма (суф.отр.гл) + ңыз (*JJ*) + дар (*KJ*), но будет правильно вывести на результат основу барма, вместо словы бар.

Грамматические суффиксы, которые являются показателями преобразования глагола табл. 1: суффиксы деепричастия; суффиксы причастия; суффиксы наклонения; суффиксы времени.

Общее количество всевозможных суффиксов составляет 26 526 единиц. Некоторые типы, приведенные в табл. 2, являются лексически и семантически допу-

Таблица 1. Индексация суффиксов

Символ	Значение	Суффиксы
Ks	Суффиксы деепричастия	-а, -е, -й, -ып, -іп, -п → (1) -ғалы, -гелі, -қалы, -келі → (2)
E	Суффиксы причастия	-ған, -ген, -қан, -кен, -атын, -етін, -йтын, -йтін → (1) -ар, -ер, -р, -с, -мақ, -мек, -бақ, -бек, -пақ, -пек → (2)
R	Суффиксы наклонения	Условное наклонение: -са, -се → (1) Желательное наклонение: -ғы, -гі, -қы, -кі → (2) Повелительное наклонение: -айын, -ейін, -йын, -йін, -іңіз, -ыңыз, -ңіз, -ңыз, -сін, -сын, -ейік, -айық, -йік, -йық, -іңдер, -ыңдар, -ңдар, -ңдер, -іңіздер, -ыңыздар, -ңіздер, -ңыздар → (3) Изъявительное наклонение: суффиксы времен
Sh	Суффиксы времени	Прошедшее время: -ып, -іп, -п → (1) -ған, -ген, -қан, -кен, -атын, -етін, -йтын, -йтін → (2) -ды, -ді, -ты, -ті → (3) Настоящее время: -ып, -іп, -п + тұр, жүр, отыр, жатыр → (4) -а, -е, -й → (5) Будущее время: -а, -е, -й → (5) -ып, -іп, -п + тұр, жүр, отыр, жатыр → (6) -ар, -ер, -р, -бақ, -бек, -пақ, -пек, -мақ, -мек

Таблица 2. Примеры образования аффиксов в словах казахского языка

Типы суффиксов и окончаний казахского языка в словах с глагольными основами	Примеры
$V + Ks1 + JJ$	бар-а-мын, көр-е-сіңдер, сөйле-й-міз, айт-ып-пыз, кел-іп-ті
$V + Ks2 \_\_ V + JJ$	бар-ғалы отыр-мын, бар-ғалы жатыр-сыздар
$V + E1 + JJ$	бар-ған-сың, бар-атын-быз
$V + E2 + TJ + CJ$	көр-ер-ің-ді, көр-мег(к → г)-ім-нің
$V + E1 + KJ + TJ + CJ$	көр-ген-дер-ім-нің, көр-етін-дер-і-не
$V + R2 + TJ \_\_ V + Ks1 + Sh3$	бар-ғы-мыз кел-е-ді

стимыми, но некоторые из них не применяются. В базу аффиксов добавлены только те, которые наиболее часто применяются.

Далее представлено, в каком порядке к основе добавляются аффиксы, для выделения правильной основы, но не были рассмотрены словообразующие суффиксы, так как они меняют основу слова и контекст значения. В основном к существительным добавляются словообразующие суффиксы.

$$W = \{N_i\} + \{E_i\},$$
$$W = \{A_i\} + \{S1_i\},$$
$$W = \{R_i\} + \{S2_i\},$$
$$W = \{V_i\} + \{S3_i\} + \{S4_i\} + \{E_i\},$$
$$S3 = \{Et, Kp\},$$
$$S4 = \{Ks1, Ks2, Sh3, Sh4, E1, R2, R1, R2, R3\},$$

где N — множество основ имен существительных и местоимений; V — множество глагольных основ; A — множество имен прилагательных; R — множество имен числительных; W — множество слов; S — множество формообразующих суффиксов, $S1$, $S2$, $S3$, $S4$; Et — суффиксы залога; Kp — усилительный вид суффиксов; Ks — суффиксы деепричастия; E — множество окончаний.

На основе выполненных исследований и разработанной системы классификации окончаний казахского языка были построены детерминированные конечные автоматы (ДКА) для различных частей речи.

На рис. 1 представлен ДКА для окончаний имен существительных казахского языка, на котором показаны возможные вариации «q_0»–«q_7» словообразования для данной части речи в казахском языке. Каждый граф описывает преобразования слова с помощью присоединения окончаний к основе и порядок присоединения, который синтаксически и семантически допустим в казахском языке.

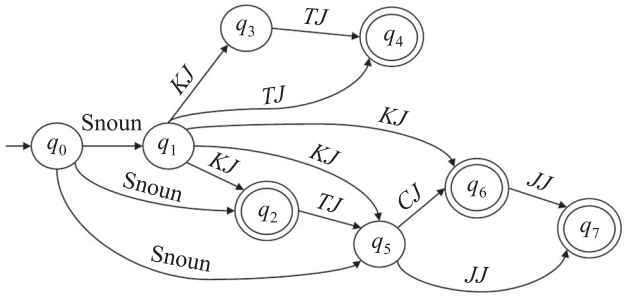

Рис. 1. Детерминированный конечный автомат для окончаний имен существительных казахского языка

Рис. 1 иллюстрирует ДКА с помощью диаграммы состояний. Для каждого состояния существует стрелка перехода, обозначающая присоединение определенного вида окончания. ДКА имеет начальное состояние (графически показывается стрелкой «из ниоткуда»), откуда начинается вычисление с переходом Snoun (основа имени существительных казахского языка), и множество (обозначаемых графически в виде двойной окружности), которые определяют с помощью образования различных вариации присоединение окончаний, успешно завершающие вычисления.

Разработка алгоритма стеминга для казахского языка

В области разработки алгоритмов лемматизации представлено достаточно много работ. Среди них есть достаточно близкие к данной работе по исследуемым языкам, и по подходу к построению самого алгоритма — алгоритм стемминга (stemming). В работе [6] рассмотрены недостатки достаточно высокого процента ошибок. В [7] предложен эффективный алгоритм стемминга для русского языка, так и для использования словаря, что повышает качество стемминга. В работе [4] предложен алгоритм лемматизации для казахского языка, в котором рассмотрена систематизация окончаний казахского языка, не обладающая полнотой. В [8] опубликован алгоритм стеминга для казахского языка. Этот алгоритм реализован на основе стеммера Портера, который охватывает только самую маленькую часть окончаний казахского языка.

Принцип предлагаемого алгоритма стеминга казахского языка, основанный на предложенной полной системе окончаний заключается в следующем.

1. В системе окончаний казахского языка все окончания разбиваются на классы по длине символов. В слове сначала ищется окончание максимальной длины для данного слова: оно будет на два символа меньше длины слова (предполагается, что основа не может меньше длины – 2). Предполагаемое окончание длины L ищется в соответствующем классе. Если окончание не находится в данном классе, то длина предполагаемого окончания уменьшается на единицу и ищется в соответствующем классе окончаний и т. д., до тех пор, пока не найдется окончание или слово будет без окончания.

2. Приняты следующие обозначения:

$L(e)^{max}$ — максимальная длина окончаний в системе окончания языка;

w — анализируемое слово;

$e(w)$ — окончание анализируемого слова;

$L(w)$ — длина анализируемого слова;

$L[e(w)]$ — предполагаемая длина окончания данного слова;

$L[e(w)]^{max}$ — максимальная длина окончания данного слова.

3. Шаги алгоритма.

Шаг 1. Определяется длина анализируемого слова $L(w)$.

Шаг 2. Определяется максимальная длина окончания анализируемого слова: $L[e(w)]^{max} = L(w) - 2$, где 2 — есть минимальная длина основы слова.

Шаг 3. $L(w) \le L(e)^{max}$, если длина слова w меньше или равно максимальной длины окончаний в системе окончаний языка, то предполагаемой длине окончания данного слова $L[e(w)]$ присваивается значение максимальной длины окончания анализируемого слова: $L[e(w)] = L[e(w)]^{max}$. Далее переход на шаг 5.

Шаг 4. Иначе: предполагаемой длине окончания данного слова $L[e(w)]$ присваивается $L(e)^{max}$: $L[e(w)] = L(e)^{max}$.

Шаг 5. Сделать выборку окончания $e(w)$ длины $L[e(w)]$ из данного слова w.

Шаг 6. Проверка $e(w)$ на совпадение с окончанием из списка окончаний длины $L[e(w)]$. Если совпадает, то определяем основу данного слова: $St(w) = w - e(w)$, т. е. из данного слова выделяется основа.

Шаг 7. Иначе: уменьшаем предполагаемую длину окончания данного слова на единицу: $L[e(w)] = L[e(w)] - 1$.

Шаг 8. Если $L[e(w)] < 1$, то слово w без окончания. Переход на шаг 9. Иначе: переход на шаг 6.

Шаг 9. Конец.

Предложенный подход представляет лексиконно свободный (lexicon free) алгоритм генерации казахского языка на основе полной системы окончаний казахского языка.

Практические результаты

Реализация алгоритма протестирована на корпусе казахского языка [9]. В заданном корпусе изначально были убраны знаки пунктуации и стоп-слова. На рис. 2 представлен интерфейс программы работы алгоритма.

На рис. 2 в левой колонке введен входной текст на казахском языке, в правой колонке — выводы результата нормализации текста. Ниже колонок расположены функциональные кнопки «Очистка окна», «Загрузка файла», «Нахождение основы слов». Практические результаты алгоритма [10, 11]:

— количество входных слов — 486 000;

— правильно определенных слов — 92 %;

— затраченное время выполнения — 2,3 с.

Для определения точности нахождения основ, полученные тексты проверены с помощью словаря казахского языка. По полученным данным можно заметить, что скорость данного алгоритма высокая и показывает хорошие результаты, но также было выявлено ряд ошибок, связанных с распознаванием основы [12] и аффиксов казахского языка. Некоторые примеры с неправильным распознаванием основы и аффиксов представлены в табл. 3.

Сұр + лау, ұзын + ырақ, жақсы + рақ, кіші + рек — здесь программой отсекаются суффиксы -лау, -ырақ, -рак/рек и выводится основа сұр, ұзын, жакс, кіш.

Жайлау, шаңырақ, тарақ, терек — здесь также программой отсекаются суффиксы -лау, -ырақ, -рак/рек и выводится основа жай, шаң, тар, тер. Но в составе этих слов они неявляются суффиксами, а являются частью основы, поэтому в этом случае выделенная программой основа слова неправильная.

Данная проблема актуальна, так как многие методы и платформы используют словари и дополнительные модули, для решении этой задачи [13–16]. В дальнейшем будут применены подходы, основанные на обучениях системы с использованием корпусов казахского языка.

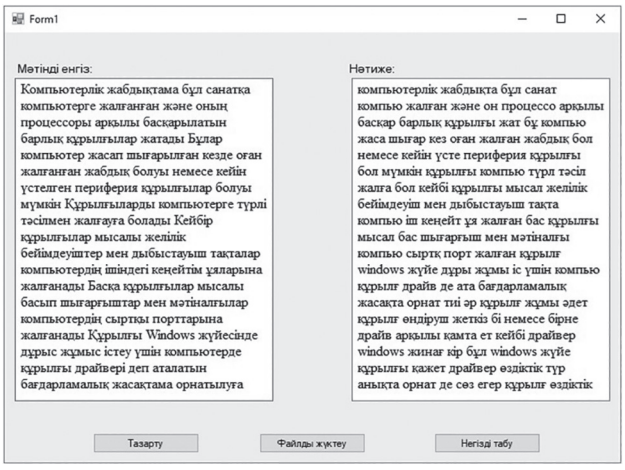

Рис. 2. Интерфейс программы применения алгоритма стемминга для казахского языка

Таблица 3. Проблемы, связанные с суффиксами степени имен прилагательных и порядковых имен числительных

Исходное слово	Выделенная программой основа слова	Правильная основа слова
Жақсырақ	Жақс	Жақсы
Кішірек	Кіш	Кіші
Алтыншы	Алт	Алты
Жетінші	Жет	Жеті

Заключение

Таким образом, исследованы современные работы систематизации окончаний казахского языка. Разработана полная система классификации окончаний и суффиксов казахского языка. Построены детерминированные конечные автоматы для существительного, прилагательного, числительного и глагола с использованием возможных вариантов добавления суффиксов и окончаний к основе для казахского языка.

Разработан алгоритм стеминга с использованием разработанной системы классификации окончаний казахского языка. По итогам анализа полученных данных алгоритм показывает хороший результат без применения словаря или иных вспомогательных модулей. Преимуществом данного подхода является скорость обработки. Полнота окончаний языка обеспечивает достаточно высокий уровень реализации. Отличительными особенностями построенного алгоритма является его достаточно легкая воспроизводимость, что позволяет, в частности, без особых трудозатрат применить его в прикладных системах обработки естественного языка.

Литература

1. Altenbek G., Wang X.-L. Kazakh segmentation system of inflectional affixes // Proc. of the CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2010). Beijing, China. 2010. P. 183–190.
2. Kessikbayeva G., Cicekli I. Rule based morphological analyzer of Kazakh language // Proc. of the 2014 Joint Meeting of SIGMORPHON and SIGFSM. Association for Computational Linguistics. Baltimore, Maryland, USA. 2014. P. 46–54. doi: 10.3115/v1/W14-2806
3. Bekmanova G., Sharipbay A., Altenbek G., Adali E., Zhetkenbay L., Kamanur U., Zulkhazhav A. A uniform morphological analyzer for the Kazakh and Turkish languages [Электронный ресурс]. URL: http://ceur-ws.org/Vol-1975/paper3.pdf (дата обращения: 10.02.2020).
4. Федотов А.М., Тусупов Д.А., Самбетбаева М.А., Еримбетова А.С., Бакиева А.М., Идрисова А.И. Модель определения нормальной формы слова для казахского языка // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2015. Т. 13. № 1. С. 107–116.
5. Тукеев У.А., Турганбаева А. Лексикон – фри стемминг для казахского языка // Материалы международной научной конференции «Информатика и прикладная математика» («Computer science and Applied Mathematics») посвященной 25-летию Независимости Республики Казахстан и 25-летию Институт информационных и вычислительных технологий. Алматы, 2016. С. 84–88.
6. Willett P. The Porter stemming algorithm: then and now // Program. 2006. V. 40. N 3. P. 219–223. doi: 10.1108/00330330610681295
7. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine [Электронный ресурс]. URL: https://www.semanticscholar.org/paper/A-Fast-Morphological-Algorithm-with-Unknown-Word-by-Segalovich/983b7014df3b7d4e82e32ba4f45f71f3879f8c96 (дата обращения: 01.03.2020).
8. Iborodikhin A. Basic snowball stemming algorith for kazakh language [Электронный ресурс]. URL: https://github.com/iborodikhin/stemmer-kaz/ (дата обращения: 27.03.2020).
9. Rakhimova D., Zhumanov Zh. Complex technology of machine translation resources extension for the Kazakh language // Studies in Computational Intelligence. 2017. V. 710. P. 297–307. doi: 10.1007/978-3-319-56660-3_26
10. Рахимова Д.Р. Разработка информационно-аналитической поисковой системы данных на казахском языке: отчет о НИР (промежуточный). № ГР 0118РК00127. Алматы, 2018. 84 с.
11. Shormakova A., Zhumanov Zh., Rakhimova D. Post-editing of words in Kazakh sentences for information retrieval // Journal of Theoretical and Applied Information Technology. 2019. V. 97. N 6. P. 1896–1908.
12. Ножов И.М. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
13. Kutuzov A., Andreev I. Texts in, Meaning out: neural language models in semantic similarity tasks for Russian // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (2015) = Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue» (2015). 2015. Т. 2. № 14. С. 133–144.
14. Kalimoldayev M.N., Koibagarov K.Ch., Pak A.A., Zharmagambetov A.S. The application of the connectionist method of semantic similarity for kazakh language // Proc. 12th International Conference

References

1. Altenbek G., Wang X.-L. Kazakh segmentation system of inflectional affixes. *Proc. of the CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2010)*, Beijing, China, 2010, pp. 183–190.
2. Kessikbayeva G., Cicekli I. Rule based morphological analyzer of kazakh language. *Proc. of the 2014 Joint Meeting of SIGMORPHON and SIGFSM. Association for Computational Linguistics*, Baltimore, Maryland, USA, 2014, pp. 46–54. doi: 10.3115/v1/W14-2806
3. Bekmanova G., Sharipbay A., Altenbek G., Adali E., Zhetkenbay L., Kamanur U., Zulkhazhav A. *A uniform morphological analyzer for the Kazakh and Turkish languages*. Available at: http://ceur-ws.org/Vol-1975/paper3.pdf (accessed: 10.02.2020).
4. Fedotov A.M., Tussupov D.A., Sambetbayeva M.A., Yerimbetova A.S., Bakieva A.M., Idrisova A.I. The implementation of the algorithm generating word forms of the Kazakh language. *Vestnik NSU. Series: Information Technologies*, 2015, vol. 13, no. 1, pp. 107–116. (in Russian)
5. Tukeev U.A., Turganbaeva A. Lexicon - free stemming for Kazakh language. *Proc. International Scientific Conference "Computer science and Applied Mathematics"*, Almaty, 2016, pp. 84–88. (in Russian)
6. Willett P. The Porter stemming algorithm: then and now. *Program*, 2006, vol. 40, no. 3, pp. 219–223. doi: 10.1108/00330330610681295
7. Segalovich I. *A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine*. Available at: https://www.semanticscholar.org/paper/A-Fast-Morphological-Algorithm-with-Unknown-Word-by-Segalovich/983b7014df3b7d4e82e32ba4f45f71f3879f8c96 (accessed: 01.03.2020).
8. Iborodikhin A. *Basic snowball stemming algorith for kazakh language*. Available at: https://github.com/iborodikhin/stemmer-kaz/ (accessed: 27.03.2020).
9. Rakhimova D., Zhumanov Zh. Complex technology of machine translation resources extension for the Kazakh language. *Studies in Computational Intelligence*, 2017, vol. 710, pp. 297–307. doi: 10.1007/978-3-319-56660-3_26
10. Rakhimova D.R. Development of information and analytical data retrieval system in Kazakh language. *Report N ГР 0118РК00127*, Almaty, 2018, 84 p. (in Russian)
11. Shormakova A., Zhumanov Zh., Rakhimova D. Post-editing of words in Kazakh sentences for information retrieval. *Journal of Theoretical and Applied Information Technology*, 2019, vol. 97, no. 6, pp. 1896–1908.
12. Nozhov I.M. *Morphological and syntactic text processing (models and programs)*. Moscow, 2003, 140 p. (in Russian)
13. Kutuzov A., Andreev I. Texts in, Meaning out: neural language models in semantic similarity tasks for Russian. *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii = Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue» (2015)*, 2015, vol. 2, no. 14, pp. 133–144.
14. Kalimoldayev M.N., Koibagarov K.Ch., Pak A.A., Zharmagambetov A.S. The application of the connectionist method of semantic similarity for kazakh language. *Proc. 12th International Conference on Electronics Computer and Computation (ICECCO 2015)*, 2015, pp. 7416906. doi: 10.1109/ICECCO.2015.7416906
15. Drakshayani B., Prasad E.V. Semantic based model for text document clustering with idioms. *International Journal of Data Engineering (IJDE)*, 2013, vol. 4, no. 1, pp. 1–13.

on Electronics Computer and Computation (ICECCO 2015). 2015. P. 7416906. doi: 10.1109/ICECCO.2015.7416906
15. Drakshayani B., Prasad E.V. Semantic based model for text document clustering with idioms // International Journal of Data Engineering (IJDE). 2013. V. 4. N 1. P. 1–13.
16. Verma R., Vuppuluri V. A New approach for idiom identification using meanings and the web // Proc. 10th International Conference on Recent Advances in Natural Language Processing (RANLP 2015). Hissar, Bulgaria. 2015. P. 681–687.

16. Verma R., Vuppuluri V. A New approach for idiom identification using meanings and the web. *Proc. 10th International Conference on Recent Advances in Natural Language Processing (RANLP 2015)*, Hissar, Bulgaria, 2015, pp. 681–687.

Авторы

Рахимова Диана Рамазановна — PhD of Computer Science, старший преподаватель, Казахский Национальный Университет имени Аль Фараби, Алматы, 050040, Казахстан; старший научный сотрудник, Институт информационных и вычислительных технологий, Алматы, 050000, Казахстан, Scopus ID: 55682794500, ORCID ID: 0000-0003-1427-198X, diana.rakhimova@kaznu.kz

Турганбаева Алия Оралбайкызы — магистр информационных систем, преподаватель, Казахский Национальный Университет имени Аль Фараби, Алматы, 050040, Казахстан, Scopus ID: 57209969959, ORCID ID: 0000-0001-9660-6928, turganbaeva.aliya@bk.ru

Authors

Diana R. Rakhimova — PhD, Senior Lecturer, Al-Farabi Kazakh National University, Almaty, 050040, Republic of Kazakhstan; Senior Scientific Researcher, Institute of Information and Computing Technologies, Almaty, 050000, Republic of Kazakhstan, Scopus ID: 55682794500, ORCID ID: 0000-0003-1427-198X, diana.rakhimova@kaznu.kz

Aliya O. Turganbayeva — Master, Lecturer, Al-Farabi Kazakh National University, Almaty, 050040, Republic of Kazakhstan, Scopus ID: 57209969959, ORCID ID: 0000-0001-9660-6928, turganbaeva.aliya@bk.ru