Стемматизация и генерация словоформ в казахском языке для систем автоматической обработки текстов

В. Б. Барахнин[1,2,*], А. М. Бакиева[2], М. Н. Бакиев[3], С. Ж. Тажибаева[3], Т. В. Батура[2,4], Л. Х. Лукпанова[1,5]

[1]Институт вычислительных технологий СО РАН, Новосибирск, Россия

[2]Новосибирский государственный университет, Россия

[3]Евразийский национальный университет им. Л. Н. Гумилева, Астана, Казахстан

[4]Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск, Россия

[5]Казахский национальный исследовательский технический университет им. К. И. Сатпаева, Алматы, Казахстан

[*]Контактный e-mail: bar@ict.nsc.ru

Предложены алгоритмы анализа и синтеза словоформ в казахском языке, основанные на принципах разбиения слов на флективные классы. Поскольку казахский язык является агглютинативным, подключать словарь словоформ для автоматизации морфологического анализа нецелесообразно. Значительно эффективнее пользоваться словарями аффиксов и наборами правил. В процессе исследования созданы словари, включающие около 2000 глагольных аффиксов и их комбинаций для 17 флективных классов и около 3500 аффиксов и их комбинаций (вариантов окончаний) для существительных и прилагательных. Некоторые сочетания аффиксов повторяются. Такой объем словарей достаточен для того, чтобы осуществлять анализ текстов любой тематической принадлежности. Предлагаемые алгоритмы могут применяться на этапе морфологического анализа в поисковых и вопросно-ответных системах, системах автореферирования, а также при построении тезаурусов и онтологий.

Ключевые слова: казахский язык, стемматизация, генерация, морфологический анализ, аффикс, флективный класс.

Введение

В связи с расширением информационного пространства появляется необходимость автоматической обработки текстов на различных языках, в частности на казахском. Казахский язык обладает богатой и сложной морфологией. Как и в других тюркских языках, слово состоит из основы, к которой присоединяются аффиксы, выражающие различные грамматические характеристики. К основе слова могут присоединяться несколько формообразующих аффиксов (иногда называемых окончаниями), при этом каждый такой аффикс выполняет присущую только ему грамматическую функцию, порядок расположения аффиксов строго определен.

В процессе тематического индексирования документа для определения его принадлежности какой-либо предметной области обычно используется некоторый набор ключевых терминов, каждый из которых обозначает какое-либо понятие из данной предметной области, причем термины встречаются в различных словоформах. Поэтому при расширенном поиске документов правильнее учитывать не словоформы, а основы слов, следовательно, необходимо создание качественного алгоритма стемматизации, т. е. выделения основы слова.

Индексация является необходимой стадией обработки текста в системах автореферирования. Такие системы позволяют получать краткое изложение содержания одного или нескольких документов. Важно, чтобы автоматически составленный реферат содержал наиболее существенные термины. Многие из систем автореферирования [1–6] не имеют поддержки для казахского языка. Описанный в данной работе алгоритм стемматизации может быть использован для индексации документов в системах автореферирования текстов на казахском языке.

Использование модуля морфологического анализа позволяет увеличить не только полноту, но и точность результата информационного поиска. Это можно объяснить тем, что в случае отсутствия процедуры стемматизации встречаются ситуации, когда в выборку попадают документы, не релевантные запросу, но содержащие совпадающие формы, в то время как в релевантных документах данные слова употребляются в другой форме. Использование частоты встречаемости основ вместо частоты встречаемости слов может позволить получить больший вес для релевантных документов и тем самым поместить их во множество отобранных. Таким образом, предлагаемый в данной статье алгоритм стемматизации может применяться в модуле морфологического анализа при поиске документов. А так как одним из способов улучшения качества поиска является использование тезаурусов [7], то очевидна важность нахождения основы для правильной работы с тезаурусом.

Морфологический анализатор играет также важную роль в вопросно-ответных системах. При поступлении вопроса в систему осуществляется его обработка, в том числе стемматизация. Согласно определенным правилам вопрос перефразируется в утвердительную форму части предложения, в котором содержится ответ. Например, вопрос “Ақтау қайда орналасқан? — где находится Актау” переформулируется в часть предложения-ответа так: “Ақтау батыста орналасқан — Актау находится на западе”. При формировании ответа необходима возможность получения различных форм слов, чтобы добиться согласования слов в генерируемом предложении [6].

В работе [8] упоминается система синтеза словоформ для русского языка, использующая словарь. Подаваемая на вход словоформа подвергается ряду преобразований на основе заложенных в систему правил, в результате чего получаются все возможные варианты исходной формы данного слова. Далее для каждого построенного таким образом варианта производится поиск его в словаре. Поскольку казахский язык является агглютинативным, использовать словарь словоформ нецелесообразно, удобнее как для стемматизации, так и для генерации пользоваться словарем аффиксов и наборами правил.

Проблемам морфологического анализа казахского языка посвящено много исследований [9–13]. Морфологические анализаторы используются в поисковых машинах для обобщения запроса пользователя. Многие алгоритмы реферирования используют частоту встречаемости слов как признак, поиск по которому дает более точные результаты, если все словоформы слова рассматривать как одно слово. Однако существующие сис-

темы автореферирования не поддерживают казахский язык, поэтому создание стемматизатора и генератора казахских словоформ является актуальной задачей. Об актуальности создания алгоритмов обработки семантической информации на казахском языке свидетельствует большое количество публикаций на эту тему (см., например, [11–13]).

В [11] описано разработанное на основе формализации морфологических правил с помощью семантических сетей программное обеспечение, названное авторами "интеллектуальным морфологическим анализатором казахского языка", в [12] тем же авторским коллективом представлена аппаратная реализация синтеза словоформ казахского языка с помощью ассоциативного запоминающего устройства, в [13] для морфологического анализа и генерации словоформ казахского языка использован подход на основе конечных автоматов.

Отличительной особенностью предложенных алгоритмов стемматизации и генерации словоформ казахского языка является использование принципа разбиения слов на флективные классы в соответствии с идеями работы [8]. С целью реализации этих алгоритмов для всех изменяемых частей речи (существительного, прилагательного, глагола) нами были описаны наборы правил сочетания аффиксов.

1. Флективные классы существительных, прилагательных и глаголов в казахском языке

В основу построения алгоритмов морфологического анализа и синтеза положено разбиение слов на классы, определяющие характер изменения буквенного состава форм слов. Эти классы условно названы морфологическими. Изменения форм слов могут носить различный характер. Они могут быть связаны с изменением как формообразующих аффиксов слова, так и его основы (что в казахском языке бывает довольно редко: так, для существительных имеется 18 исключений, для глаголов — 352).

Морфологические классы слов делятся на два вида [8]: основоизменительные, характеризующие систему изменения слов, и флективные. Флективные классы изменяемых слов выделялись на основе анализа их синтаксической функции и систем падежных, личных и родовых окончаний. Классы неизменяемых слов выделялись только по синтаксическому принципу. В данной статье более подробно рассмотрим флективные классы глаголов в казахском языке, поскольку флективные классы существительных и прилагательных были описаны в работе [9]. В таблице приведены примеры окончаний флективных классов для некоторых аффиксов времен и лиц. Следуя правилам грамматики казахского языка [14], для глаголов мы установили 17 флективных классов, зависящих от окончания основы слова: типа последней (предпоследней, если слово заканчивается на *у, ю*) гласной основы и последней буквы основы вообще. Отметим, что в казахском языке гласные звуки подразделяются на твердые (*а, о, ұ, ы*) и мягкие (*ә, е, ө, ү, і*).

Выделены следующие варианты чередования гласных и согласных в основе слова:

1) твердый гласный (*а, о, ұ, ы*), основа оканчивается на гласный (кроме *ю*) — тара*у*;
2) мягкий гласный (*ә, е, ө, ү, і*), основа оканчивается на гласный (кроме *ю*) — төле*у*;
3) твердый гласный, основа оканчивается на согласные *б, г* — таб*у*, бағ*у*;
4) мягкий гласный, основа оканчивается на согласные *б, г* — теб*у*, тіг*у*;

Флективные классы с набором окончаний в 3-м лице

Номер флективного класса	Отрицание	Прошедшее время, 3-е лицо	Субъективное прошедшее время, 3-е лицо	Результативное прошедшее время, 3-е лицо	Конкретное настоящее время, 3-е лицо	Переходное время, 3-е лицо	Переходное прошедшее время, 3-е лицо	Будущее продленное время, 3-е лицо	Будущее время намерения, 3-е лицо
1	ма	ды	пты	ған	п тұр	йды	йтын	р	мақ
2	ме	ді	пті	ген	п тұр	йді	йтін	р	мек
3	па	ты	ыпты	қан	ып тұр	ады	атын	ар	пақ
4	пе	ті	іпті	кен	іп тұр	еді	етін	ер	пек
5	ба	ды	ыпты	ған	ып тұр	ады	атын	ар	бақ
6	бе	ді	іпті	ген	іп тұр	еді	етін	ер	бек
7	ма	ды	пты	ған	ып тұр	ады	атын	ар	мақ
8	ме	ді	пті	ген	іп тұр	еді	етін	ер	мек
9	ба	ды	пты	ған	ып тұр	ады	атын	ар	бақ
10	бе	ді	пті	ген	іп тұр	еді	етін	ер	бек
11	ыма	ыды	ыпты	ған	ып тұр	иды	итын	ар	ымақ
12	іме	іді	іпті	ген	іп тұр	иді	итін	ер	імек
13	па	ты	ыпты	қан	ып тұр	ады	атын	ар	пақ
14	пе	ті	іпті	кен	іп тұр	еді	етін	ер	пек
15	ма	йды	йыпты	ған	ып тұр	яды	ятын	яр	ймақ
16	ме	йді	йіпті	ген	іп тұр	еді	етін	ер	ймек
17	ма	ды	ыпты	ған	ып тұр	ады	атын	ар	мақ

5) твердый гласный, основа оканчивается на согласный *з* — жаз*у*;
6) мягкий гласный, основа оканчивается на согласный *з* — жүз*у*;
7) твердый гласный, основа оканчивается на согласные *р, л* — бар*у*, сал*у*;
8) мягкий гласный, основа оканчивается на согласные *р, л* — күл*у*, көр*у*;
9) твердый гласный, основа оканчивается на согласные *м, н, ң* — дам*у*, тан*у*, тоң*у*;
10) мягкий гласный, основа оканчивается на согласные *м, н, ң* — кем*у*, түн*у*, жең*у*;
11) твердый гласный, основа оканчивается на согласные *ж, д* — алж*у*, аңд*у*;
12) мягкий гласный, основа оканчивается на согласные *ж, д* — ренж*у*, дегд*у*;
13) твердый гласный, основа оканчивается на глухой согласный — жарас*у*;
14) мягкий гласный, основа оканчивается на глухой согласный — күрес*у*;
15) твердый гласный, основа оканчивается на *ю* — жа*ю*;
16) мягкий гласный, основа оканчивается на *ю* — тү*ю*;
17) твердый гласный, основа оканчивается на *у* — жу*у*.

В таблице приведены примеры окончаний слов флективных классов для некоторых аффиксов времен и лиц.

2. Алгоритмы генерации и стемматизации словоформ

Пошаговое описание алгоритма генерации существительных приведено в статье [9]. Аналогичным образом выполняется генерация словоформ для прилагательных и глаголов. Единственное отличие состоит в первом шаге. На вход подается существительное или прилагательное в именительном падеже и единственном числе; глагол — в форме инфинитива.

2.1. Алгоритм генерации словоформ глаголов

Для глаголов имеются следующие виды окончаний (в скобках каждый вид окончаний обозначен заглавной латинской буквой):

1) окончание отрицания P_1;
2) окончание времени P_2;
3) личное окончание P_3.

Возможны следующие комбинации окончаний:

1) окончание времени P_2 (например, “ген” — ‘бер-ген’);
2) окончание времени + личное окончание (P_2P_3), например “ген + сің” — ‘істе-ген-сің’;
3) окончание отрицания + окончание времени (P_1P_2), например “пе + ді” — ‘кет-пе-ді’;
4) окончание отрицания + окончание времени + личное окончание ($P_1P_2P_3$), например “ма + ды + ңыз” — ‘оқы-ма-ды-ңыз’.

Еще раз отметим, что порядок присоединения аффиксов строго фиксирован и обусловлен флективным классом.

2.2. Алгоритм стемматизации словоформ глаголов

Приведем алгоритм стемматизации глаголов (алгоритм стемматизации существительных и прилагательных описан в работе [10]). В его основе лежит алгоритм Портера [15].

В зависимости от выполнения условий констатируется, получена ли основа слова или требуется отсечение аффикса. Алгоритм получения основы Q состоит из следующих этапов.

1. На вход поступает любая словоформа.
2. Начиная с последней буквы слова, происходит поиск по списку аффиксов.
3. Если данный аффикс найден, то он отсекается. Оставшаяся часть слова после отсечения всех аффиксов считается основой.

Каждое слово z запишем в виде $z = x_0 \wedge x_1 \wedge \cdots \wedge x_k$ как конкатенацию частей слова $x_0, x_1, \ldots, x_k$. Если слово имеет вид $z = x \wedge y$ (далее будем использовать обозначение $y(x)$), то $y = z \backslash x$.

Алгоритм соответствует следующему набору правил A (далее $x = z \backslash x_{k+1-i}$, где i — номер этапа).

Случай 1.

$$z = Q \wedge P_2, \quad A = \{\ P_2(x) \to Q.$$

Случай 2.

$$z = Q \wedge P_2 \wedge P_3, \quad A = \begin{cases} P_3(x) \to P_2(Q), \\ P_2(x) \to Q. \end{cases}$$

Случай 3.

$$z = Q \wedge P_1 \wedge P_2, \quad A = \begin{cases} P_2(x) \to P_1(Q), \\ P_1(x) \to Q. \end{cases}$$

Случай 4.

$$z = Q \wedge P_1 \wedge P_2 \wedge P_3, \quad A = \begin{cases} P_3(x) \to P_2(P_1(Q)), \\ P_2(x) \to P_1(Q), \\ P_1(x) \to Q. \end{cases}$$

Если при вводе слова допущена опечатка (например, если во введенном слове нет гласных), то определить подходящий флективный класс не удается и генератор выдает сообщение об ошибке.

Следует отметить, что предложенные алгоритмы стемматизации и генерации применимы к простым формам глаголов. Более сложные формы глаголов, состоящие из 2–4 слов, планируется рассмотреть в дальнейшем. Однако в научно-технических текстах сложные глаголы практически не используются.

3. Реализация и тестирование предложенных алгоритмов

Ранее было разработано веб-приложение для генерации словоформ существительных и прилагательных, описанных в работе [9]. К этому приложению добавлена реализация изложенных выше алгоритмов генерации словоформ глаголов и стемматизации существительных, прилагательных и глаголов. Приложение находится в открытом доступе в сети Интернет [16]. Рекомендуемые браузеры Google Chrome, Mozilla FireFox. Веб-приложение демонстрирует принципиальные возможности системы. Хотя веб-интерфейс несколько замедляет работу (словоформы заданного слова генерируются примерно за 25 с), реальное применение разработанного авторами программного обеспечения в непосредственной интеграции с лингвистическими программными системами позволяет получать результаты быстрее: время генерации всех словоформ конкретного слова 1 с. Предполагается обращаться к настоящей версии, а не к веб-приложению.

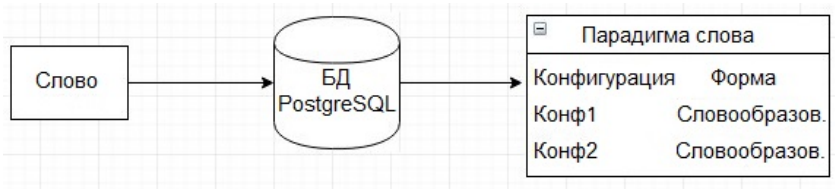
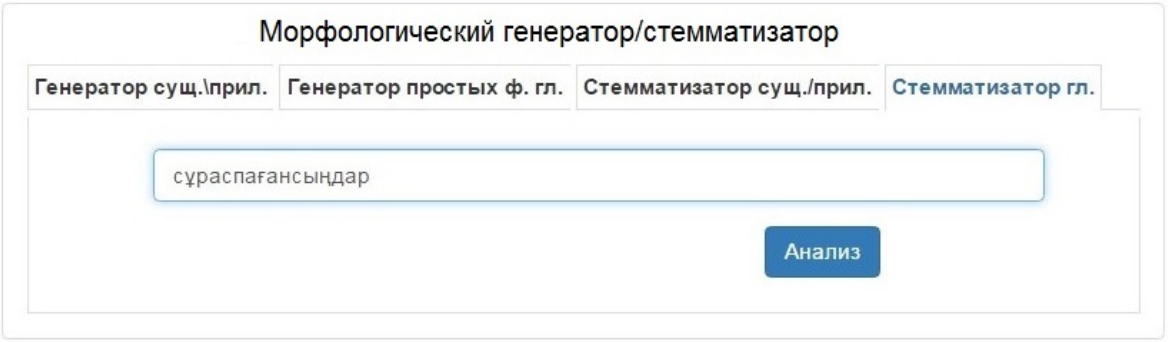


Рис. 1. Пример архитектуры системы генерации форм слова



Конфигурация словообразования	Форма
stem	сұрас

Рис. 2. Пример стемматизатора словоформ глаголов

Конфигурация словообразования	Форма
Отрицание	болмау
Вопрос	болды ма
Результативно прош.вр.	болған
Прошедшее время	болды
Переходное время	болады
Будущее продолжительное время	болар
Переходное прошедшее время	болатын
Конкретное наст. вр	болып тұру
Будущее время намерения	болмақ
Давнопрошедшее время	болыпты
Результативно прош.вр. + Личное окон. 1л ед.ч.	болғанмын
Результативно прош.вр. + Личное окон. 2л ед.ч.	болғансың
Результативно прош.вр. + Личное окон. 2ув.л ед.ч.	болғансыз
Результативно прош.вр. + Личное окон. 1л мн.ч.	болғанбыз
Результативно прош.вр. + Личное окон. 2л мн.ч.	болғансыңдар

Рис. 3. Пример генерации словоформ глаголов

Модуль генерации и модуль стемматизации реализованы на языке Python с использованием библиотек psycopg2, collections. Словари хранятся в базе данных PostgreSQL. На рис. 1 приведена архитектура программного обеспечения.

При тестировании на словах, принадлежащих различным частям речи, не было обнаружено ошибок, что позволяет судить о корректности предложенных алгоритмов.

На рис. 2 показан результат работы созданного стемматизатора, на рис. 3 — список словоформ, полученных с помощью созданного алгоритма генерации.

Заключение

Изложены алгоритмы стемматизации и генерации глаголов, что в совокупности с результатами из работ [9, 10] полностью решает задачу анализа и синтеза словоформ для научно-технических текстов на казахском языке. В процессе исследования для существительных и прилагательных выделено 14 флективных классов, а для глаголов — 17. Созданы словари, включающие более 5500 аффиксов и их комбинаций (с учетом повторений комбинаций для различных грамматических форм). Количественный объем словарей достаточен для того, чтобы осуществлять анализ текстов любой тематической принадлежности. Система дополнена словарем исключений, содержащим 18 существительных и 352 глагола, в которых при словоизменении изменяется основа. При тестировании на словах, принадлежащих различным частям речи, не было обнаружено ошибок, что позволяет судить о корректности предложенных алгоритмов.

Разработанные алгоритмы могут применяться на этапе морфологического анализа в поисковых и вопросно-ответных системах, системах автореферирования, а также при построении тезаурусов и онтологий.

Благодарности. Работа выполнена при частичной поддержке Президентской программы “Ведущие научные школы РФ” (грант № 7214.2016.9).

Авторы выражают искреннюю признательность аспиранту ИВТ СО РАН Илье Сергеевичу Пастушкову за помощь при размещении программного приложения в сети Интернет.

Список литературы / References

[1] **Тревгода С.А.** Методы и алгоритмы автоматического реферирования текста на основе анализа функциональных отношений: Автореф. дис. ... канд. техн. наук. СПб., 2009. 18 с.
Trevgoda, S.A. Methods and algorithms for automatic text summarization based on analyzing functional relationship: Abstr. of dis. for the degree of candidate of technical sci. St. Petersburg, 2009. 18 p. (In Russ.)

[2] **Гридина Е.А.** Анализ алгоритмов автоматического реферирования текста // Вост.-Европ. журн. передовых технологий. 2011. Т. 3, № 2(51). С. 36–38.
Gridina, E.A. Analysis of algorithms for automatic text summarization // Eastern-Europ. J. of Enterprise Technologies. 2011. Vol. 3, No. 2(51). P. 36–38. (In Russ.)

[3] **Хан У., Мани И.** Системы автоматического реферирования. Адрес доступа: http://www.osp.ru/os/2000/12/067_print.htm (дата обращения: 12.03.2015)
Han, U., Mani, I. Systems of automatic summarization. Available at: http://www.osp.ru/os/2000/12/067_print.htm (accessed: 12.03.2015) (In Russ.)

[4] **Гинкул А.С.** Сравнительный анализ существующих систем автоматического реферирования текста // Політ. сучасні проблеми науки. Киев, 2012. С. 255.
Ginkul, A.S. A comparative analysis of the existing systems of automatic text summarization // Polit. Modern Problems of Sci. Kiev, 2012. P. 255. Available at: http://jrnl.nau.edu.ua/index.php/Fly/article/view/2598. (In Russ.)

[5] **Анно Е.Н.** Система морфологического анализа с синтезом словоформ // Семиотика и информатика. 1978. Вып. 10. С. 168–187.
Anno, E.N. The system of morphological analysis with the synthesis of word forms // Semiotics and Informatics. 1978. Iss. 10. C. 168–187. (In Russ.)

[6] **Monz, C.** Document retrieval in the context of question answering // Proc. of the 25th Europ. Conf. on Inform. Retrieval Res. (ECIR-03) / F. Sebastiani (Ed.). Lecture Notes in Comput. Sci. 2003. Vol. 2633. P. 571–579.

[7] **Шокин Ю.И., Федотов А.М., Барахнин В.Б.** Проблемы поиска информации. Новосибирск: Наука, 2010. 196 с.
Shokin, Yu.I., Fedotov, A.M., Barakhnin, V.B. Information retrieval problems. Novosibirsk: Nauka, 2010. 196 p. (In Russ.)

[8] **Белоногов Г.Г., Зеленков Ю.Г.** Алгоритм автоматического анализа русских слов // Вопр. информ. теории и практики. 1985. № 53. С. 62–93.
Belonogov, G.G., Zelenkov, Yu.G. Algorithm for automatic analysis of Russian words // Theor. and Pract. Iss. of Journalism. 1985. No. 53. P. 62–93. (In Russ.)

[9] **Барахнин В.Б., Лукпанова Л.Х., Соловьев А.А.** Алгоритм построения словоформ с использованием флективных классов для систем морфологического анализа казахского языка // Вестн. НГУ. Информ. технологии. 2014. Т. 12, вып. 2. С. 25–31.
Barakhnin, V.B., Lukpanova, L.Kh., Solovyev, A.A. The algoritm for constructing wordforms using inflexional classes for systems of Kazakh language morfological analysis // Novosibirsk State Univ. J. of Inform. Technologies. 2014. Vol. 12, iss. 2. P. 25–31. (In Russ.)

[10] **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 // Вестн. НГУ. Информ. технологии. 2015. Т. 13, вып. 1. С. 107–116.
Fedotov, A.M., Tusupov, D.A., Sambetbayeva, M.A., Yerimbetova, A.S., Bakiyeva, A.M., Idrisova, A.I. The implementation of the algorithm generating word forms of the Kazakh language // Novosibirsk State Univ. J. of Inform. Technologies. 2015. Vol. 13, iss. 1. P. 107–116. (In Russ.)

[11] **Шарипбаев А.А., Бекманова Г.Т., Ергеш Б.Ж., Бурибаева А.К., Карабалаева М.Х.** Интеллектуальный морфологический анализатор, основанный на семантических сетях // Матер. Междунар. науч.-техн. конф. “Открытые семантические технологии проектирования интеллектуальных систем” (OSTIS-2012). Минск, БГУИР, 16–18 февраля 2012. С. 397–400.
Sharipbaev, A.A., Bekmanova, G.T., Ergesh, B.Zh., Buribaeva, A.K., Karabalaeva, M.Kh. Intelligent morphological analyzer, based on semantic networks // Proc. of the Intern. Sci.-Techn. Conf. “Open Semantic Intelligent Systems Design Technology” (OSTIS-2012). Minsk, BGUIR, February of 16–18. 2012. P. 397–400. (In Russ.)

[12] **Бурибаева А.К., Шарипбаев А.А., Бекманова Г.Т., Ергеш Б.Ж., Карабалаева М.Х.** Аппаратная реализация синтеза словоформ казахского языка с помощью ассоциативной памяти // Вестн. Евраз. нац. ун-та им. Л.Н. Гумилева. 2012. Спец. выпуск. С. 180–183.

Buribaeva, A.K., Sharipbaev, A.A., Bekmanova, G.T., Ergesh, B.Zh., Karabalaeva, M.Kh. Hardware implementation for the synthesis of word forms in the Kazakh language using associative memory // Bulletin of the Euras. Nat. Univ. L.N. Gumilev. 2012. Special issue. P. 180–183. (In Russ.)

[13] **Заурбеков Д.Л., Кайракбай Б.М.** Построение конечного преобразователя для морфологического анализа и генерации словоформ казахского языка // Materiały VIII Międzynar. nauk.-prakt. konf. "Wschodnie partnerstwo–2012". Przemyśl, 07–15 września. Vol. 8. Filologiczne nauki. Przemyśl: Nauka i studia, 2012. S. 30–39.
Zaurbekov, D.L., Kayrakbay, B.M. Construction of the final drive for morphological analysis and generation of word forms in the Kazakh language // Proc. of VIII Intern. Sci.-Pract. Conf. "Eastern Partnership–2012". Przemyśl, 07–15 Sept. Vol. 8. Filologiczne Nauki. Przemyśl: Nauka i Studia, 2012. P. 30–39. (In Russ.)

[14] **Валяева Т.** Грамматика казахского языка. Адрес доступа: http://kaz-tili.kz (дата обращения: 20.01.2017)
Valyaeva, T. The grammar of the Kazakh language. Available at: http://kaz-tili.kz (accessed: 20.01.2017) (In Russ.)

[15] **Porter, M.F.** An algorithm for suffix stripping // Program. 1980. Vol. 14, No. 3. P. 130–137.

[16] **Бакиева А.М.** Программа генерации словоформ казахского языка. Адрес доступа: http://db4.sbras.ru/morpher
Bakieva, A.M. Program generation of word forms of the Kazakh language. Available at: http://db4.sbras.ru/morpher (In Russ.)

Поступила в редакцию 6 марта 2017 г.,
с доработки — 13 июня 2017 г.

Stemming and generation of word forms in automatic text processing systems in the Kazakh language

Barakhnin, Vladimir B.,[1,2,*], Bakiyeva, Aigerim M.[2], Bakiyev, Murat N.[3], Tazhibayeva, Saule Zh.[3], Batura, Tatiana V.[4], Lukpanova, Lyazzat Kh.[1,5]

[1]Institute of Computational Technologies SB RAS, Novosibirsk, 630090, Russia
[2]Novosibirsk State University, Novosibirsk, 630090, Russia
[3]L. N. Gumilyov Eurasian National University, Astana, 010008, Kazakhstan
[4]A. P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk, 630090, Russia
[5]K. I. Satpayev Kazakh National Research Technical University, Almaty, 050013, Kazakhstan
[*]Corresponding author: Barakhnin, Vladimir B., e-mail: bar@ict.nsc.ru

Purpose. Currently there is an urgent need for automatic processing of texts in the Kazakh language. Morphological analysis in the process of automatic text processing allows increasing both the completeness and the accuracy of the result of information retrieval. Since the Kazakh language is agglutinative, it is impractical to use the dictionary of word forms for the automation of morphological analysis. It is much more effective to use affix dictionaries and sets of rules. Algorithms for synthesizing and analyzing word forms of the Kazakh language are proposed in this article.

Methodology. A distinctive feature of the proposed algorithms for stemming and generation of word forms of the Kazakh language is the use of the principle of words

splitting into inflectional classes. To implement these algorithms for all changeable parts of speech (noun, adjective, verb), we described the sets of affix combination rules.

Findings. During the research the dictionary was developed. It includes about 2000 verbal affixes and their combinations for the 17 inflectional classes and about 3500 affixes and their combinations (variants of endings) for nouns and adjectives. Some combinations of affixes are repeated. The system is supplemented with an exception dictionary, including 18 nouns and 352 verbs, in which the word forms are formed by changing the stem. Such a volume of the dictionaries is sufficient to perform text analysis of any themes. The generation module and the stemming module are implemented in Python using libraries: psycopg2, collections. The dictionaries are stored in the database PostgreSQL.

Originality. We tested the software application on words belonging to different parts of speech, and found no errors, which makes it possible to judge the correctness of the proposed algorithms. The proposed algorithms can be applied at the stage of morphological analysis in the search engines, summarization systems and question-answer systems, as well as in the construction of thesauri and ontologies.

Keywords: Kazakh language, stemming, generation, morphological analysis, affixes, inflectional classes.

Acknowledgements. Work is executed with partial support of the Presidential programme "Leading scientific schools of RF" (grant No. 7214.2016.9).

Received 6 March 2017
Received in revised form 13 June 2017