Алгоритм синтеза словоформ казахского языка с использованием флективных классов

© В.Б. Барахнин © Л.Х. Лукпанова © А.А. Соловьев
Институт вычислительных технологий СО РАН,
Новосибирский государственный университет,
Казахский национальный технический университет им. К.И. Сатпаева
bar@ict.nsc.ru lyazzat.lukpanova@mail.ru salerat@gmail.com

Аннотация

В статье рассматриваются принципы разбиения существительных казахского языка на флективные классы и построение основанного на этом разбиении алгоритма синтеза словоформ.

Работа выполнена при частичной поддержке РФФИ (проекты 12-07-00472, 13-07-00258), президентской программы «Ведущие научные школы РФ» (грант НШ 5006.2014.9) и интеграционных проектов СО РАН.

1 Введение

При создании информационных систем, предназначенных для обработки семантической информации, представленной на том или ином естественном языке, возникает необходимость генерации всех словоформ изменяемых слов данного языка. Разумеется, объем словаря любого естественного языка подразумевает требование автоматизации процесса генерации словоформ. Более того, поскольку словарь естественного языка постоянно пополняется новыми словами (прежде всего – научными терминами), постольку задачу генерации словоформ нельзя решить «раз и навсегда», вследствие чего необходим простой и понятный алгоритм синтеза словоформ, который допускает возможность достаточно простой программной реализации.

Для русского языка алгоритмы морфологического анализа и синтеза слов, основанные на разбиении всех слов на морфологические классы, каждый из которых характеризуется определенным типом изменения буквенного состава форм слов, входящих в этот класс, подробно изложены в монографии [1]. Морфологические классы русского языка делятся на два вида: основоизменительные и флективные, при этом для каждой изменяемой части речи существует, вообще говоря, свое разбиение на морфологические классы. Флективные классы строятся основе анализа синтаксической функции слов и систем их падежных, родовых и личных окончаний. Поскольку в русском языке не существует зависимости типа изменения окончаний от буквенного состава основы слова (так, к разным флективным классам относятся схожие по буквенному составу слова *боль* и *соболь*; *волос, голос* и *колос* и т.п.), постольку флективные классы определяются их типичными представителями, при этом число классов весьма велико: например, для существительных в [1] выделено 66 флективных классов. Ввиду этого процесс отнесения слова к тому или ному флективному классу может быть автоматизирован лишь частично: эксперт должен вручную выбрать флективный класс (чтобы сузить число объектов-альтернатив до приемлемого для возможности эффективного ручного выбора, эксперту следует предварительно определить грамматические характеристики слова).

За время, прошедшее с момента публикации монографии [1], изложенные в ней алгоритмы убедительно продемонстрированы свою эффективность, в том числе и при их реализации на базе современных информационных технологий (см., например, работы [2, 3], в которых представлена модификация алгоритма синтеза словоформ, применяющаяся при пополнении базового лексического словаря программной библиотеки, используемой в системах информационного обеспечения научной деятельности).

Отметим, что альтернативные алгоритмы решения рассматриваемой задачи демонстрируют намного бóльшую сложность при меньшей точности. Так, в работе [4] классы словоформ определялись без учета теоретических изысканий из монографии [1]: путем непосредственного анализа типов окончаний (не в морфологическом, а в «обычном» смысле) слов, что привело к появлению более 10 тысяч классов для существительных (к одному классу в [4] были отнесены слова, у

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

начальных форм которых совпадают 3 последние буквы). Такое количество классов делает алгоритм трудновоспроизводимым. Однако даже столь детальное разбиение не способно дать абсолютно точное различение слов по типу склонения, как показывает приведенный выше пример), к тому же «эмпирический» характер разбиения может вызвать определенные вопросы относительно полноты описания классов.

В настоящее время возникает необходимость решения подобной задачи для государственных языков некоторых государств СНГ, в которых начинают активно развиваться интернет-технологии, в частности для казахского (отметим, что на сегодняшний день ни Google, ни Yandex не предоставляют возможность автоматического перевода с казахского языка или на казахский язык). Об актуальности создания алгоритмов обработки семантической информации на казахском языке свидетельствует большое количество публикаций на эту тему (см., например, [5–7]).

В [5] описан разработанный в Евразийском национальном университете им. Л.Н.Гумилева интеллектуальный морфологический анализатор казахского языка на основе формализации морфологических правил с помощью семантических сетей, в [6] тем же авторским коллективом представлена аппаратная реализация синтеза словоформ казахского языка с помощью ассоциативного запоминающего устройства, наконец, в [7] сотрудниками Казахского национального технического университета им. К.И. Сатпаева для морфологического анализа и генерации словоформ казахского языка использован подход на основе конечных автоматов.

Перечисленные работы отличает, с одной стороны, детальное изложение особенностей реализации алгоритмов (вплоть до заполнения конкретных регистров ассоциативного запоминающего устройства в [6] и фрагментов программного кода в [7]), а с другой стороны – почти полное отсутствие описания лежащей в основе алгоритмов теоретической базы, относящейся к области компьютерной лингвистики. В итоге алгоритмы, описываемые в перечисленных работах, трудновоспроизводимы, поскольку за деталями их реализации практически не видна их суть, хотя выполнение требования воспроизводимости алгоритмов генерации словоформ весьма важно, поскольку создатели информационных систем (прежде всего, научной тематики) постоянно вынуждены пополнять лексические словари этих систем словоформами новых терминов. Необходимость требования воспроизводимости алгоритмов могла бы быть снята наличием общедоступного веб-приложения для генерации словоформ произвольных слов, но такие приложения ни для казахского, ни для русского языка нам неизвестны (мы не случайно сделали оговорку относительно произвольности слов, так как, например, веб-приложение для морфологического анализа слов русского языка [8], созданное на основе Грамматического словаря А.А. Зализняка, выдает словоформы лишь тех слов, которые содержатся в словаре программы).

В настоящей статье рассматриваются принципы разбиения существительных казахского языка на флективные классы и построение основанного на этом разбиении алгоритма синтеза словоформ (мы ограничили подробный анализ лишь существительными, поскольку, как показано в [3], в качестве «новых» слов при пополнении лексического словаря практически всегда выступают существительные или прилагательные, однако в казахском языке прилагательные, выступающие в роли определений, не приобретают окончаний, а изменение прилагательных, выступающих в роли существительных, не отличается от изменения существительных).

2 Разбиение существительных казахского языка на флективные классы

Казахский язык относится к классу агглютинативных языков [9], т.е. словоформы в нем образуются путем добавлению к корню аффиксов (суффиксов и окончаний). Например, для имен существительных к корню сначала добавляется суффикс, а далее: окончание множественного числа, притяжательное окончание, падежное окончание и, наконец, личное окончание предикативности. Принципиальным отличием морфологии казахского языка от морфологии русского является наличие в казахском языке (как и в других тюркских языках) закона сингармонизма, в соответствии с которым аффиксы слова полностью определяются звуковым составом его основы.

На основании анализа работ [9, 10] можно выделить следующие основные правила казахского языка:

– После твердого слога следует твердое окончание, после мягкого слога следует мягкое окончание.

– Порядок следования окончаний таков: первым окончанием идет окончание множественного число, вторым – притяжательное окончание и только третьим – окончание какого-либо падежа. Т.е. ***слово*** + **число** + **притяжательность** + **падеж.** Например: адамдарымызға (нашим людям): ***адам*** + **дар** + **ымыз** + **ға** / множественное число + 1 лицо, мн. число + дательно-направительный падеж (идущее последним личное окончание предикативности в данной работе не рассматривается, поскольку для научной лексики указанная словоформа нехарактерна).

Мягкость и твердость слов в казахском языке определяется наличием определенной гласной в последнем слоге слова. Например, слово твердое, когда присутствуют гласные **а, о, ұ, ы, я**, а мягкое, когда присутствую гласные **ә, ө, ү, і, е**. Твердость или мягкость слов коррелирует также с наличием

некоторых согласных: слово твердое, если в нем присутствуют согласные **қ** и **ғ**, и мягкое, если присутствуют **к** и **г**.

Каждое следующее окончание зависит от предыдущего по нескольким параметрам:

– По твердости: если последний слог слова твердый, то каждое следующее окончание будет твердым, так как твердость очередного окончания зависит от предыдущего. Таким образом, если слово твердое, то все окончания твердые, если мягкое, то мягкие.

– По последней букве окончания: каждое последующее окончание зависит от последней буквы предыдущего окончания.

Исследуя структурированные правила присоединения окончаний, приведенные в [10], мы установили для существительных казахского языка 14 флективных классов:

1) твердое, оканчивается на гласную (кроме **у**);
2) твердое, оканчивается на **б, в, г, д**;
3) твердое, оканчивается на **ж, з**;
4) твердое, оканчивается на **л**;
5) твердое, оканчивается на **м, н, ң**;
6) твердое, оканчивается на **р, у, й**;
7) твердое, оканчивается на глухую согласную;
8) мягкое, оканчивается на гласную (кроме **у**);
9) мягкое, оканчивается на **б, в, г, д**;
10) мягкое, оканчивается на **ж, з**;
11) мягкое, оканчивается на **л**;
12) мягкое, оканчивается на **м, н, ң**;
13) мягкое, оканчивается на **р, у, й**;
14) мягкое, оканчивается на глухую согласную.

Перечисленное разбиение слов на флективные классы полностью и без пересечений покрывает все возможные варианты слов казахского языка, что означает полное корректное решение поставленной задачи (можно отметить, что некоторые подварианты не реализуются: например, как отмечено выше, твердые слова не могут оканчиваться на букву **г**, однако мы указали и это сочетание, чтобы не нарушалась формальная полнота покрытия).

В таблице 1 приведены окончания флективных классов для падежей и множественного числа.

Подчеркнем, что в отличие от русского языка, для которого процесс отнесения слова к определенному флективному классу может быть автоматизирован лишь частично, принадлежность слова казахского языка к тому или иному флективному классу жестко детерминирован его буквенным составом. Это обстоятельство заметно облегчает реализацию алгоритма, так и его практическое использование.

Таблица 1. Флективные классы с набором окончаний

№ флективн. класса	Исходн. падеж	Местн. падеж	Дат.-направит. падеж	Родит. падеж	Винит. падеж	Творит. падеж	Мн. число
1 / 8	дан / ден		ға / ге	ның / нің	ны / ні	мен	лар / лер
2 / 9	тан / тен	та / те	қа / ке	тың / тің	ты / ті	пен	тар / тер
3 /10		да / де	ға / ге	дың / дің	ды / ді	бен	дар / дер
4 /11			ға / ге	дың / дің	ды / ді	мен	дар / дер
5 /12	нан / нен	да / де	ға / ге	ның / нің	ды / ді	мен	дар / дер
6 /13		да / де	ға / ге	дың / дің	ды / ді	мен	лар / лер
7 /14	тан / тен	та / те	қа / ке	тың / тің	ты / ті	пен	тар / тер

3 Алгоритм генерации словоформ

Для существительных возможны следующие комбинации окончаний (в скобках указан условный код данной комбинации):

1) число (1),
2) притяжательность (2),
3) падеж (3),
4) число + притяжательность (12),
5) число + падеж (13),
6) притяжательность + падеж (23),
7) число + притяжательность + падеж (123).

Так как в казахском языке следующее окончание зависит от предыдущего, то для генерации форм окончаний типов 12 и 13 требуются сгенерированные варианты форм окончаний типа 1, для форм типа 23 – формы типа 2 и т.п. Значит, окончания генерируются в следующем порядке:

1) 1, 2, 3;
2) 12, 13, 23;
3) 123.

Окончания типов 1, 2, 3 идут непосредственно сразу после основы слова, то есть они непосредственно зависят от флективного класса данного слова. При генерации составных окончаний добавление того или иного очередного окончания определяется буквенным составом предыдущего, однако весь процесс добавления окончаний однозначно определяется принадлежностью основы слова к тому или иному флективному классу.

В итоге нами были сгенерированы все варианты окончаний существительных – для 14 флективных классов получилось около 3500 вариантов окончаний.

Поскольку на вход алгоритма подается слово в именительном падеже единственного числа, не имеющее окончаний, постольку возникает задача нормализации словоформ «новых» слов, когда они употреблены в тексте не в начальной форме. К сожалению, полностью автоматическое решение этой задачи вряд ли возможно, так как последний слог основы некоторых слов (в том числе заимствованных) может совпадать с одним из окончаний казахского языка, и для выяснения того, какая именно из этих альтернатив реализована в данном конкретном случае, требуется, как минимум, синтаксический анализ предложения, в котором встретилось «новое» слово.

Таким образом, алгоритм генерации словоформ состоит из следующих этапов.

1. На вход подается слово в именительном падеже единственного числа.

2. Слово разбивается на слоги. Определяется твердость / мягкость слова.

3. По последней букве слова с учетом твердости / мягкости определяется флективный класс.

4. С использованием таблицы окончаний флективных классов генерируются все словоформы данного слова.

На выходе получается таблица, содержащая все возможные словоформы данного слова.

4 Практическая реализация и тестирование алгоритма

Нами было разработано веб-приложение для генерации словоформ существительных казахского языка, размещенное в открытом доступе сети Интернет [11]. Пример работы приложения показан на рис. 1 (для краткости приведена только часть словоформ).

Введите слово в именительном падеже (пример: адам) адам

Сгенерировать

Конфигурация слово образования	**Форма слова**
Исходный падеж	адамнан
Местный падеж	адамда
Дательно-направит. падеж	адамға
Родительный падеж	адамның
Винительный падеж	адамды
Творительный падеж	адаммен
Множественное число	адамдар
Отрицание, вопрос	адамба
1 лицо, ед.число	адамым
1 лицо, мн.число	адамымыз
2 лицо, ед. число	адамың
2 лицо, мн. число	адамың
2 лицо, (ув) ед. число	адамыңыз
2 лицо, (ув) мн. число	адамыңыз
3 лицо, ед. число	адамы
3 лицо, мн. число	адамы
Притяжательные окончания 2-ая форма	адамдікі
Множественное число + Исходный падеж	адамдардан
Множественное число + Местный падеж	адамдарда
Множественное число + Дательно-направит. падеж	адамдарға
Множественное число + Родительный падеж	адамдардың
Множественное число + Винительный падеж	адамдарды

Рисунок 1. Пример работы приложения

Была проведена проверка правильности генерации словоформ. Для этого с сайта грамматики казахского языка [10] случайным образом выбиралось слово, а также определялась одна из его словоформ, которая будет использована для тестирования. После этого с помощью приложения [11] генерировались все словоформы данного слова, и выяснялось, правильно ли сгенерирована словоформа, выбранная для тестирования (один из авторов статьи является носителем казахского языка).

В итоге для 400 произвольно выбранных словоформ было получено 100% правильно сгенерированных, из чего следует, что алгоритм работает корректно.

5 Заключение

В настоящей статье построено разбиение существительных казахского языка на флективные классы и изложен основанный на этом разбиении алгоритма синтеза словоформ. Отличительными особенностями построенного алгоритма являются его понятность и достаточно легкая воспроизводимость, что позволяет, в частности, без особых трудозатрат применить его для генерации словоформ других изменяемых частей речи казахского языка, прежде всего, глаголов. Было разработано веб-приложение для генерации словоформ существительных казахского языка, размещенное в открытом доступе сети Интернет [11]. Тестирование показало корректность его работы.

Литература

[1] Г.Г. Белоногов, А.П. Новоселов. Автоматизация процессов накопления, поиска и обобщения информации. М.: Наука, 1979.

[2] В.Б. Барахнин, А.А. Куперштох. Алгоритм координатного индексирования электронных научных документов // Труды международной конференции «Вычислительные и информационные технологии в науке, технике и образовании». Казахстан, Павлодар, 20–22 сентября 2006 г. Т. I. С. 228–232.

[3] Ю.И. Шокин, А.М. Федотов, В.Б. Барахнин. Проблемы поиска информации. Новосибирск: Наука, 2010.

[4] Е.А. Каневский. Некоторые вопросы пополнения морфологического словаря терминами предметной области // Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. Аксаково, 2001. Т. 2. С. 156–160.

[5] А.А. Шарипбаев, Г.Т. Бекманова, Б.Ж. Ергеш, А.К. Бурибаева, М.Х. Карабалаева. Интеллектуальный морфологический анализатор, основанный на семантических сетях // Материалы международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем» (OSTIS-2012). Минск, БГУИР, 16–18 февраля 2012 г. С. 397–400.

[6] А.К. Бурибаева, А.А. Шарипбаев, А.А. Бекманова А.А., Б.Ж. Ергеш, М.Х. Карабалаева. Аппаратная реализация синтеза словоформ казахского языка с помощью ассоциативной памяти // Вестник Евразийского национального университета им. Л.Н. Гумилева, 2012. Специальный выпуск. С.180–183.

[7] Д.Л. Заурбеков, Б.М. Кайракбай. Построение конечного преобразователя для морфологического анализа и генерации словоформ казахского языка // Materiaiy VIII Mikdzynarodowej naukowi-praktycznej konferencji «Wschodnie partnerstwo – 2012». Przemyśl, 07-15 września 2012 r. Vol. 8. Filologiczne nauki. Przemyśl: Nauka i studia. S. 30–39.

[8] Морфологический анализатор. http://starling.rinet.ru/cgi-bin/morphque.cgi?encoding=win

[9] Қазақ грамматикасы. Фонетика, сөзжасам, морфология, синтаксис. Астана: Астана полиграфия, 2002.

[10] Грамматика казахского языка. http://kaz-tili.kz/

[11] Программа генерации словоформ казахского языка. http://my.ict.nsc.ru/~salerat/kaz/

The Algorithm for Synthesis of the Wordforms of Kazakh Language Using Inflexional Classes

V.B. Barakhnin, L.Kh. Lukpanona, A.A. Solovyev

The article considers the principles of Kazakh nouns partition into inflexional classes and development of wordform synthesis algorithm based on such partition.