3rd International Conference on Computer Processing
in Turkic Languages (TURKLANG 2015)

Автоматные модели анализа морфологии и полнота системы окончаний казахского языка

Уалшер Тукеев [1]

Казахский Национальный университет им. аль-Фараби, кафедра Информационных систем, пр.Аль-Фараби 71, Алматы, Казахстан

Abstract

In this paper we investigate the question of the completeness of the endings in a source language in automaton models of morphological analysis of an agglutinative languages, in particular, the Kazakh language. A complete system of the endings of the Kazakh language is created. A comparative analysis with previously built the endings system of Kazakh by Bektayev is made. Completeness of endings of analyzed languages ensures that all the words of the input text will be correctly analyzed at the morphological level.

Keywords: morphology; analyze; endings; language; Kazakh; completeness;

1. Введение

Вопрос морфологического анализа являются важными в процессе обработки естественных языков. Определяющим подходом морфологического анализа является двухуровневая морфология, предложенная (Koskenniemi, 1983), реализуемая через использование конечных преобразователей. Системы морфологического анализа языков, основанные на двухуровневой морфологии, используют аппарат теории автоматов, а именно, аппарат теории и методологии конечных преобразователей (finite state transducers-FST). Имеются публикации в области использования технологии двухуровневой морфологии

1

* Corresponding author. *E-mail address:* ualsher.tukeyev@kaznu.kz

и FST для агглютинативных языков (Oflazer, 1994; Washington et al., 2014; Kairakbay and Zaurbekov, 2013; Kessikbayeva and Cicekli, 2014). В работе (Тукеев и др., 2013) авторы использовали класс FST, а именно, табличные многозначные отображения для анализа морфологии инфлективных языков, относящиеся к недетерминированным FST.

В применении технологии FST для инфлективных языков существенным является определение множества возможных окончаний слов анализируемых языков. В случае построения FST в виде графового представления, множество возможных окончаний слов будет представляться множеством возможных путей на графовой модели. В случае представления в виде табличных многозначных отображений, множество возможных окончаний слов будет представляться множеством строк таблицы многозначных отображений. В любом из этих представлении анализа морфологии весьма существенным является вопрос: является множество возможных окончаний слов анализируемого языка полным? Если, по тем или иным причинам, множество возможных окончаний слов анализируемого языка будет неполным, то возможна ситуация, что такое слово будет проанализировано неправильно или неуспешно. В связи с вышеизложенным проблема определения полноты множества возможных окончаний слов анализируемого языка является весьма важной.

Рассмотрим полноту множества окончаний на модели анализа морфологии казахского языка в виде следующих многозначных отображений:

$$F_s: X_s \rightarrow Y_s \text{ (для исходного языка)},$$
$$F_t: Y_t \rightarrow Z_t \text{ (для целевого языка)},$$

где X_s – окончания исходного языка,

Y_s – грамматические характеристики слова исходного языка,

Y_t - грамматические характеристики слова целевого языка,

Z_t - окончания целевого языка.

В данной системе отображений для обеспечения корректности преобразований любого слова пары языков машинного перевода необходимо, чтобы множество окончаний целевого и (или) исходного языка было полным. Полнота множества окончаний исходного языка весьма важным является для морфологического анализа предложений исходного языка, так как является гарантией того, что любое слово будет проанализировано на предмет его грамматической (лексической) характеристики.

В данной работе рассмотрим полноту системы окончаний казахского языка.

Так как полнота системы окончаний одного языка в паре машинного перевода определяет косвенно и в целом полноту системы преобразований на морфологическом уровне с одного языка на другой язык пары, то это является важным вопросом для всей системы машинного перевода.

2. Полнота системы окончаний казахского языка

Рассмотрим систему окончаний слов казахского языка двух классов: окончания к именным основам (существительные, прилагательные, числительные) и окончания к глагольным основам (глаголы, причастия, деепричастия, наклонения, залоги).

2.1. Система окончаний казахского языка к именным основам

Система окончаний к именным основам слов казахского языка имеет четыре типа:
- окончания множественного числа (обозначим через K),
- притяжательные окончания (обозначим через T),
- падежные окончания (обозначим через C),
- личные окончания (обозначим через J),
- основу(stem) обозначим через S.

Рассмотрим всевозможные варианты размещений типов окончаний: из одного типа, из двух типов, из трех типов и из четырех типов. Число размещений определяется формулой:

$A_n^k = n!/(n-k)!$.

Тогда, количество размещений будет определяться следующим образом:

$A_4^1 = 4!/(4-1)!=4$,
$A_4^2 = 4!/(4-2)!=12$,
$A_4^3 = 4!/(4-3)!=24$,
$A_4^4 = 4!/(4-4)!=24$.

Всего возможных размещений 64.

Рассмотрим какие из них семантически допустимы.

Размещения по одному типу окончания (K, T, C, J) являются все семантически допустимыми по определению.

Размещения по два типа окончаний могут быть следующие:

KT, TC, CJ, JK
KC, TJ, CT, JT
KJ, TK, CK, JC.

Анализ семантики размещений двух типов окончаний показывает, что выделенные жирным шрифтом размещения являются допустимыми (**KT, TC, CJ, KC, TJ, KJ**), а остальные размещения относим к недопустимым. Например, TK – после притяжательных окончаний окончания множественного числа не используются, CK – после падежных окончаний не принято ставить окончание множественного числа, JC- после личных окончаний не принято ставить падежные окончания, CT – после падежных окончаний не ставятся притяжательные окончания, JT- после личных окончаний не ставятся

притяжательные окончания. Отнесем к недопустимым и JK- после личных окончаний окончания множественного числа, так как этот тип размещения покрывается окончаниями множественного числа личных окончаний.

Вообще, типы окончаний **T** и **J** являются окончаниями определения зависимости субъектов, объектов, действий. В словах с именными основами для **TJ** двойное определение зависимости возможно для случаев различения субстанций(субъектов, объектов, действий): *апа-ң-мын (апа-ң относится к другому субъекту, а личное окончание –мын определяет зависимость к говорящему)*. В случае TJ двойное определение зависимости к одной и той субстанции запрещается, например: *апа-м-мын* не говорят.

Необходимо отметить, что тип окончаний **CJ** имеет ограничения по падежам ілік (родительный - genitive) и табыс(винительный - accusative).

Итак, количество допустимых (правильных) размещений из двух типов окончаний будет равно 6.

Размещения окончаний из трех типов будут следующие:

KTC, **KTJ**, **TCJ**, TCK, CJK, CJT, JKT, JKC

KCJ, KCT, TJK, TJC, CTK, CTJ, JTK, JTC

KJT, KJC, TKC, TKJ, CKT, CKJ, JCK, JCT.

Определение допустимых размещений окончаний из трех типов сделаем по правилу:

если в размещении из трех типов есть недопустимые размещения из двух типов, то это размещение – недопустимо.

Тогда, допустимых размещений окончаний из трех типов будет 4 (**KTC**, **KTJ**, **TCJ, KCJ** выделено жирным).

Размещения окончаний из четырех типов будут следующие:

KTJC, TKJC, CKTJ, JKTC

KTCJ, TKCJ, CKJT, JKCT

KJTC, TJKC, CTKJ, JTKC

KJCT, TJCK, CTJK, JTCK

KCTJ, TCJK, CJKT, JCKT

KCJT, TCKJ, CJTK, JCTK

Определение допустимых размещений окончаний из четырех типов сделаем по правилу:

если в размещении из четырех типов есть недопустимые размещения из двух типов, то это размещение – недопустимо.

Тогда, допустимых размещений окончаний из четырех типов будет 1 (**KTCJ** выделено жирным).

Итого, допустимых размещений из одного типа – 4, из двух типов - 6, из трех типов – 4, из четырех типов – 1.

Итак, суммарное число типов допустимых размещений в словах с именными основами – 15.

Ниже в таблице 1. представлены 15 типов окончаний слов казахского языка с именными основами с примерами и соответствующей грамматической структурой этих типов в русском языке с примерами. При описании соответствующей грамматической структуры на русском языке используются теги Penn Treebank (http://www.clips.ua.ac.be) и плюс PPRN – притяжательные местоимения, PPRNPL - притяжательные местоимения множественного числа.

Таблица1. Типы окончаний слов казахского языка с именными основами с примерами и соответствующей грамматической структурой этих типов в русском языке

Типы окончаний казахского языка в словах с именными основами	Примеры на казахском языке	Адекватная грамматическая структура на русском языке	Примеры на русском языке
S-K	тәте-лер	N+PL	Тет-и
S-T	тәте –м	PPRN N	моя тетя
S-J	тәте –мін	PRN - N	Я-тетя
S-C	тәте –ге	IN N	К тете
S-K-T	тәте-лер-ім	PPRN N+PL	мои тети
S-K-J	тәте-лер-міз	PRN - N+PL	Мы-тети
S-K-C	Тәте-лер-ге	IN N+PL	К тетям
S-T-J	Тәте-м-сіз	PRN - PPRN N	Вы-моя тетя
S-T-C	Тәте-м-ге	IN PPRN N	К моей тете
S-J-K	Тәте-сің-дер	PRN - N	Вы-тети
S-C-J	Тәте-ден-сің	PRN IN N	Вы - от тети
S-K-T-J	Тәте-лер-ім-сіндер	PRN - PPRNPL N+PL	Вы – мои тети
S-K-T-C	Тәте-лер-ім-ге	IN PPRNPL N+PL	К моим тетям
S-K-C-J	Тәте-лер-ге-мін	PRN IN N+PL	Я к тетям
S-T-C-J	Тәте-ң-нен-біз	PRN - IN PPRN N	Мы- от твоей тети
S-K-T-C-J	Тәте-лер-ің-ге-міз	PRN - IN PPRNPL N+PL	Мы-к твоим тетям

2.2. Система окончаний казахского языка к глагольным основам

Система окончаний казахского языка к глагольным основам включает следующие виды:

- система окончаний глаголов;
- система окончаний причастий;
- система окончаний деепричастий;
- система окончаний наклонений;
- система окончаний залогов.

Система окончаний к глагольным основам (глаголы) включают следующие типы:

- времена (8 времен),

- лицо (3 вида),
- отрицание.

Тогда, количество возможных типов окончаний глаголов будет -25 .

Система окончаний к глагольным основам причастия включают следующие типы:

- окончания причастия (обозначим R),
- окончания множественного числа (обозначим К),
- окончания притяжательные (Т),
- падежные окончания (обозначим С)
- личные окончания (обозначим J).

Тогда, возможные варианты типов окончаний (тип окончаний причастия для всех вариантов одинаков) будут:

- с одним типом окончаний:

RK, RT, RC, RJ;

- с двумя типами окончаний:

RKT, RTC, RCJ, RJK
RKC, RTJ**,** R*CT, RJT*
RKJ, R*TK, RCK, RJC;*

- с тремя типами окончаний:

RKTC, RTCJ, RCJK, RJKT
RKTJ, RTCK, RCJT, RJKC
RKCJ, RTJK, RCTK, RJTK
RKCT, RTJC, RCTJ, RJTC
RKJT, RTKC, RCKT, RJCK
RKJC, RTKJ, RCKJ, RJCT;

- с четыремя типами окончаний:

RKTJC, RTKJC, RCKTJ, RJKTC
RKTCJ, RTKCJ, RCKJT, RJKCT
RKJTC, RTJKC, RCTKJ, RJTKC
RKJCT, RTJCK, RCTJK, RJTCK
RKCTJ, RTCJK, RCJKT, RJCKT
RKCJT, RTCKJ, RCJTK, RJCTK.

Рассмотрим семантическую допустимость вариантов окончаний.

Все варианты окончаний причастий по одному типу окончаний являются семантически допустимыми.

Анализ семантики размещений двух типов окончаний причастий показывает, что выделенные жирным шрифтом размещения являются допустимыми, а остальные размещения относим к недопустимым. Допустимые варианты окончаний причастий такие же, как в системе окончаний с именными основами, но из них для причастий являются недопустимым вариант RTJ, так как последовательность «окончание причастия-притяжательные окончания»

для причастий во всех случаях означает персонофицированное действие с глагольной основой. А персонофицированное действие не может второй раз персонофицироваться личным окончанием. Например, *бар-ған-ым(мой приход, my coming)*, однако нельзя сказать *бар-ған-ым-сың(ты – мой приход)*, так как действие *(бар-ған-ым)* не персонофицируется, т.е. действие не может представиться субъектом.

Аналогично, окончания RTCJ и RKTCJ не имеют ограничений по двум типам окончаний, т.е. возможные пары окончаний внутри этих типов окончаний являюся допустимыми, но они нарушают предыдущее правило «действие не может представиться субъектом». Например, для RTCJ: *бар-ған-ым-ға-мын*, где «*бар-ған-ым-ға»(к моему приходу – to my coming)* – склонение действия, что не может представиться субъектом. Для RKTCJ: *бар-ған-дар-ың-нан-біз*, где «*бар-ған-дар-ың-нан*»(*от твоих приходов – from yours coming*) – склонение действий, что не может представиться субъектами.

Таким образом, количество типов окончаний причастий составляет – 11.

Рассмотрим типы окончаний деепричастий. Они представляются окончаниями переходного будущего времени, за которыми следует личные окончания: PJ, где P – базовое окончание деепричастия, J – личные окончания. Для данного класса выделим только следующие базовые окончания: *-ғаны, -гелі, -қалы, -келі.* Таким образом, считаем, что количество типов окончаний деепричастия -1.

Рассмотрим окончания наклонений, а именно, условного, повелительного, желательного. Окончания изъявительного наклонения совпадают с окончаниями глаголов в настоящем, прошлом и будущем.

Тип окончаний склонений аналогичен предыдущему, т.е. базовые окончания наклонений, за которыми следуют личные окончания. Таким образом, будем считать, что имеются три типа окончаний наклонений: условного, повелительного, желательного.

Типы окончаний залогов, а именно, возвратного, страдательного,совместного и принудительного, также определяются по предыдущей схеме: базовые окончания залогов за которыми следуют личные окончания. Соответсвенно, типов окончаний залогов будет – 4.

Итак, общее количество типов окончаний слов с глагольными основами будет 48.

Итого, общее количество окончаний с именными основами плюс общее количество типов окончаний слов с глагольными основами будет равно 63.

Следующей задачей является по полученным типам окончаний определить формы окончаний и их количество. Это сделать несложно так как для каждого типа части речи имеются соответствующие правила. В данном направлении автором построены конечные множества окончаний для всех основных частей

речи казахского языка. Так, для частей речи с именными основами количество окончаний равно 862, а количество окончаний частей речи с глагольными основами составляет: глаголы – 432, причастия- 1588, деепричастия- 48, наклонения – 230, залоги- 80. Итого, 3240 всего окончаний.

3. Сравнительный анализ разработанной системы окончаний казахского языка с моделью Бектаева

Данная работа является развитием модели Бектаева (Бектаев, 1999) для применения в области машинного перевода. В модели Бектаева определено множество окончаний казахского языка в количестве 753 окончаний. Бектаев в своей модели предложил также алгоритм использования разработанной им системы окончаний казахского языка в связи их с грамматическими характеристиками для правильного и точного перевода (немашинного) на другой язык. В модели Бектаева определены 15 типов окончаний слов с именными основами. Предлагаемая модель также имеет 15 типов окончаний, однако отличается двумя типами: в модели Бектаева используются типы JK и TJK, которые нами отнесены к окончаниями множественного числа личных окончаний. В модели Бектаева для слов с глагольными основами используются три типа окончаний с причастиями, в то время как в предлагаемой модели – 11 типа окончаний с причастиями, один тип деепричастия, три типа окончаний наклонений: условного, повелительного, желательного, четыре типа окончаний залогов, а именно, возвратного, страдательного, совместного и принудительного. В общем итоге, количество окончаний в предлагаемой модели 3240 против 753 окончаний модели Бектаева.

4. Заключение и дальнейшие работы

В работе сделана попытка построения полной системы окончаний казахского языка, что будет являться основанием для полноты системы морфологического анализа текстов на казахском языке. Так как казахский язык относится к агглютинативной группе языков, то вопрос полноты системы окончаний других языков данной группы может быть исследован аналогично. Данный подход, по мнению автора, позволит повысить качество морфологического анализа, соответственно, и качество машинного перевода. В качестве дальнейших работ планируется использовать результаты данного исследования в разработке систем машинного перевода, проводимых автором и его исследовательской группой.

Данные исследования проводятся в рамках грантового финансирования 0749/ГФ4 Министерства образования и науки Республики Казахстан.

References

Koskenniemi, K. (1983). *Two-level morphology: A general computational model of word-form recognition and production*. Tech. rep. Publication No. 11. Department of General Linguistics. University of Helsinki.

Gurenko, V.V. (2013). .*Intoduction to automata theory*. Electronic handbook. – M.: MGTU, -pp. 62(на русском языке).

Oflazer, K. (1994). *Two-level description of Turkish morphology*, Literary and Linguistic Computing Volume9, Issue2. 137-148.

Washington, J. N., Salimzyanov, I., Tyers, F.M. (2014). *Finite-state morphological transducers for three Kypchak languages*. Proceedings of the 9th Conference on Language Resources and Evaluation.

Kairakbay, B.M., Zaurbekov, D. L. (2013). Finite *State Approach to the Kazakh Nominal Paradigm.*. Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing. St Andrews–Sctotland. 108–112.

Kessikbayeva, G., Cicekli, I. (2014). Rule *Based Morphological Analyzer of Kazakh Language*. Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, Baltimore, Maryland USA . 46–54.

Тукеев, У.А., Рахимова, Д.Р., Байсылбаева, К., Умирбеков, Н., Оразов, Б., Абақан, М., Кызырканова, С., (2013). Көпмағыналық бейнелеу кесте тәсілі негізінде орыс тілінен қазақ тіліне машиналық аудармасының морфологиялық анализбен синтезін құру. Түркі тілдерін компьютерлік өңдеу. Бірінші халықаралық конференция: Еңбектері/ Астана: Л.Н.Гумилев атындағы ЕҰУ баспасы, 182-191.

Bektayev, K. (1999). *Big Kazakh-Russian and Russian-Kazakh dictionary*, Almaty, “Altyn Kazyna”. pp.704 (in Kazakh, Russian).