МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

Писменото наследство и информационните технологии [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.

ISBN 978–954–9787–25–2

Алгоритм работы морфологического парсера калмыцкого языка

В. В. Куканова, А. Ю. Каджиев

Калмыцкий язык, Национальный корпус калмыцкого языка, морфологический анализатор, автоматическая обработка текстов, алгоритм программы

An Algorithm for a Morphological Parser of the Kalmyk Language

Viktoria Kukanova, Arasha Kadzhiev

This article is devoted to the description of an algorithm for a morphological parser of the Kalmyk language for automatic processing of texts. The work on the creation and development of the National Corpus of the Kalmyk Language began at the end of 2010 and now includes 17 million of tokens. The parser analyzes about 90 percent of text including 11 % of homonym variants for morphological annotation. Our aim is to increase the figure of non-ambiguous analysis.

Корпусная лингвистика активно развивается в отечественной науке, в особенности в регионах России. Например, появились корпусы татарского, башкирского, бурятского и других языков, т. е. тех, которые не носят международного статуса. Калмыцкий институт гуманитарных исследований РАН приступил к созданию корпуса калмыцкого языка в конце 2010 г. На данный момент НККЯ включает 17 млн словоупотреблений и состоит из художественных произведений (поэтических и прозаических) и газетных текстов разного характера (интервью, статей, объявлений, репортажей и т. д.).

Целью данной статьи является описание алгоритма работы морфологического анализатора для калмыцкого языка, необходимого для автоматической обработки текстов. Токены представляют собой объект анализа морфологического анализатора, назначение которого заключается в осуществлении стемминга, лемматизации и собственно морфологического анализа.

Калмыцкий язык по своей структуре принадлежит к числу агглютинативных, что влияет на алгоритм работы морфологического анализатора. Анализ строился на словарном подходе: использовался электронный словарь калмыцкого языка с грамматической информацией, где у каждого слова выявлены стемы и указано, какой парадигме то или иное слово принадлежит.

Грамматический словарь основывался на словнике Калмыцко-русского словаря — единственной лексикографической работе академического характера [КРС 1977]. Отдельно был создан словарь аффиксов, который постоянно пополняется новыми единицами. Аффиксы имеют свою грамматическую характеристику, которая используется в работе анализатора. Таблица аффиксов включает в себя следующую информацию: 1) аффикс и его уникальный ключ; 2) грамматическая

информация (граммема); 3) список частей речи, к которым может примыкать данный аффикс (см. таблицу 1).

Табл. 1. Фрагмент таблицы аффиксов

№	*Аффикс*	*Часть речи*	*Граммема*
1.	*-а*	N; ADJ; NUM	GEN
2.	*-снлә*	PTCPL	PST.COM
3.	*-агдна*	V	CAUS1.PASS.PRES
4.	*-лдҗәхинь*	PTCPL	RECP.DUR2.FUT.ACC2.3POSS
5.	*-мүдәснь*	N; ADJ	PL.ABL.3POSS
6.	*-навдн*	V	PRES.1PLPER

На входе имеется словарь основ слов разных частей речи — местоимений, числительных, послелогов, частиц, союзов, идеофонов (звукоподражательных слов) и др. Таблица основ состоит из 1) основы и ее уникального ключа; 2) номера стемы; 3) части речи; 4) леммы (начальной формы); 5) перевода; 6) характеристики с лексико-грамматической и семантической точек зрения; 7) парадигмы (поле не носит обязательного характера, поскольку не все части речи изменяются) (см. таблицу 2).

Табл. 2. Фрагмент таблицы основ

№	*Основа*	*Часть речи*	*Характеристика*
1.	*хан*	N	lemma:хан; r:concr; t:hum; translation: хан; царь; монарх; N-4(S*H)/д
2.	*хаан*	N	lemma:хан; r:concr; t:hum; translation: хан; царь; монарх; N-4(S*H)/д
3.	*хаа*	N	lemma:хан; r:concr; t:hum; translation: хан; царь; монарх; N-4(S*H)/д
4.	*ха*	N	lemma:хан; r:concr; t:hum; translation: хан; царь; монарх; N-4(S*H)/д

Морфологический анализатор непосредственно обращается к основе, ее номеру, лемме и парадигме. Остальная информация (перевод, характеристика) дублируется при записи леммы. Существует еще одна таблица, где прописаны словоизменительные типы и правила словоизменения внутри этого типа. Данная таблица необходима для того, чтобы снять множественные вероятностные разборы.

Алгоритм работы морфологического анализатора состоит из следующих шагов.

1. Парсер обращается к словарю основ неизменяемых частей речи, а также словоформ, имеющих нестандартное склонение (что проявляется в наличии супплетивных основ), находит совпадения, после чего записывает вероятностный результат морфологического анализа.

2. Анализатор обращается к словарю аффиксов, ищет их в токенах от конца, начиная с самых длинных и заканчивая самыми короткими. Программа строит гипотезы, какие аффиксы можно вычленить в данном слове. Их может быть несколько.

Например, в словоформе *теңгсин* выделяются следующие вероятностные аффиксы: *-син* (N/ADJ; PL.GEN), *-ин* (N/ADJ; GEN), *-н* (CONV; MOD), в *негдгч* — *-дгч* (NUM; COL.NOM/COL.ACC2), *-гч* (CONV; MOM), *-ч* (CONV; IPFV), в *өгчәнә* — *-чәнә* (V; DUR2.PRES), *-нә* (V; PRES).

3. Анализатор отрезает от конца токена найденные цепочки аффиксов, в результате чего получаются стемы. Как правило, их количество равняется количеству найденных аффиксов. Несовпадение числа аффиксов и стемов происходит по причине наличия омонимичных основ.

Например, от указанных выше токенов анализатор отрезает найденные аффиксы, в итоге получаются стемы: *теңг~~сын~~* → *теңг-*, *теңгс~~ин~~* → *теңгс-*, *теңгси~~н~~* → *теңгси-*; *нег~~дгч~~* → *нег-*, *негд~~гч~~* → *негд-*, *негдг~~ч~~* → *негдг-*; *өгч~~әнә~~* → *өг-*, *өгчә~~нә~~* → *өгчә-*.

4. Полученные вероятностные стемы сравниваются с основами, которые даны в словаре. Цель этой операции заключается в том, чтобы выявить те стемы, которые прописаны в словаре. Одновременно из дальнейшего анализа устраняются те аффиксы, которые были ошибочно выделены на первом этапе.

Анализатор находит в словаре основ следующие последовательности: *теңг-* (N), *теңгс-* (N), *нег-* (NUM), *негд-* (V), *өг-* (V) — и устраняет из дальнейшего анализа стемы *теңгси-*, *негдг-*, *өгчә-*, которые были найдены при ошибочном выделении аффиксов *-н*, *-ч*, *-нә*.

5. Если анализатор успешно находит основы, формируется список гипотетических основ слова. Затем происходит сравнение информации, прописанной в словаре основ и словаре аффиксов, для того, чтобы уменьшить количество вероятностных разборов. Информация по принадлежности основы к определенной части речи должна совпадать с информацией о части речи в словаре основ.

На данном этапе при сравнении выявляется совпадение частеречных характеристик основ и аффиксов или его отсутствие:

теңг- (N)	=	*-син* (N/ADJ; PL.GEN)
теңгс- (N)	=	*-ин* (N/ADJ; GEN)
нег- (NUM)	=	*-дгч* (NUM; COL.NOM/COL.ACC2)
негд- (V)	=[1]	*-гч* (CONV; MOM)
өг- (V)	=	*-чәнә* (V; DUR2.PRES)

6. На этом этапе подключается таблица парадигм. Анализатор, обращаясь к ней, генерирует форму в соответствии с прописанными правилами словоизменения. Потом сравниваются токен и синтезированная форма:

[1] Поскольку конвербы образованы от глагольных основ, то в данном случае основа личной формы глагола и деепричастия совпадает.

теңг- (N)	*-син (N/ADJ; PL.GEN)*	≠	*N-2CC(SH)/д*	*Stem 2 + -дин (PL.GEN)*
теңгс- (N)	*-ин (N/ADJ; GEN)*	=	*N-C(СШ, С)/муд*	*Stem + ин (GEN)*
нег- (NUM)	*-дгч (NUM; COL.NOM/COL.ACC2)*	=	*Num-2CC(SH)/д*	*Stem + -дгч (COL.NOM/COL.ACC2)*
негд- (V)	*-гч (CONV; MOM)*	=	*V-C (СДӐ)*	
өг- (V)	*-чәнә (V; DUR2.PRES)*	=	*V-C (СГ)*	*Stem + -чәнә (DUR2.PRES)*

7. После данной операции токену приписывается вероятностная лемма, парадигма, грамматическая и семантическая характеристика, которая дана в словаре аффиксов.

8. Последний этап основан на бессловарном способе определения грамматической характеристики. Мы решили ввести данный этап для того, чтобы увеличить количество вероятностных разборов. В калмыцком языке встречаются такие аффиксы, которые не омонимичны концам слов и не совпадают с ними. Анализатор находит такие уникальные концы слов и строит гипотезы об основе и части речи, которой может принадлежать стем (таблица 3).

Табл. 3. Фрагмент итоговой таблицы аффиксов

Аффикс	*Часть речи*	*Граммема*
-лһчклавидн	V	CAUS1.COMPL.REM.1PLPER
-мүдтәһиньӊ	N; NUM; ADJ; PTCPL	PL.ASSOC.ACC1.3POSS

В 2013 г. в целях тестирования системы управления содержимым сайта и процедур поисковых запросов была запущена тестовая версия Национального корпуса калмыцкого языка (www.kalmcorpora.ru) без морфологической и семантической разметки, хотя данный тип аннотации был осуществлен в закрытой базе данных. Материал с разметкой будет опубликован после доработки программного кода анализатора и увеличения количества разборов с 70 % до 90 %, в том числе и множественных. В настоящее время около 20 % текста имеют множественные вероятностные варианты автоматического анализа. У 9 % токенов отсутствует результат морфологического анализа ввиду того, что в словаре основ отсутствуют их стемы (в основном это слова из русского языка, не вошедшие в Калмыцко-русский словарь под ред. Б. Д. Муниева [КРС 1977], а также русские собственные имена). Алгоритм работы морфологического анализатора, возможно, будет изменяться и дополняться, но его принципы останутся без изменений.

Литература

КРС 1977 — *Калмыцко-русский словарь* / под ред. Б. Д. Муниева. М.: Русский язык, 1977. 768 с.