УДК 81'44

АВТОМАТИЧЕСКИЙ МОРФОЛОГИЧЕСКИЙ АНАЛИЗ ДЛЯ КОРПУСОВ ТЮРКСКИХ ЯЗЫКОВ

© А.В.Дыбо, А.В.Шеймович

В статье описываются принципы работы автоматического морфологического анализатора для тюркских языков. Выделяются его основные компоненты: грамматический словарь языка; порядковая модель словоформы (набор позиций в словоформе и морфонологических представлений аффиксов для этих позиций); правила сочетаемости аффиксов в пределах словоформы и двухуровневые фонетические правила выбора алломорфов конкретного аффикса. В основе работы парсера лежит алгоритм анализа, разработанный Ф.Крыловым на базе системы StarLing. В качестве иллюстрации приводится версия морфологического анализатора для древнетюркского языка. Работа ведется в рамках масштабного проекта по созданию корпусных ресурсов языков народов России, в частности корпусов на малых тюркских языках РФ.

Ключевые слова: корпус языка, морфологический парсер, морфологический анализ, словоизменение, компьютерная модель словоформы, система StarLing.

0. Уже несколько лет благодаря поддержке Президиума РАН развиваются исследования по новой программе: корпусная лингвистика. Отдельное направление этой программы – корпуса языков народов России. А.В.Дыбо является одним из координаторов этого направления и руководит проектом корпусов миноритарных тюркских языков. В рамках этого проекта предполагается делать параллельные корпуса (все тексты обеспечены русским переводом) с морфологической разметкой. Аналогичный проект в рамках того же направления развивают И.В.Кормушин и И.А.Невская; это корпус древнетюркского языка, продолжающий и развивающий работу М.Эрдала и И.А.Невской в рамках корпуса VATEC [1]. Все материалы по этим двум проектам будут представлены в свободном доступе в Интернете.

Автоматические морфологические анализаторы для тюркских языков, используемые в больших программных продуктах, таких как Abbyy Finereader, Abbyy Lingvo и многие другие, основаны на тех же принципах, что и анализаторы, разработанные для флективных индоевропейских языков типа русского или немецкого. Каждая лемма в словаре получает индекс типа парадигмы, который отсылает к списку образцов парадигм. Поскольку во флективных языках эти парадигмы сами по себе невелики, а число их, наоборот, велико, анализатор строит полную парадигму для каждого слова и затем сравнивает попавшуюся в тексте словоформу с этой парадигмой. Для тюркских языков используется тот же метод, что позволяет не подвергать изменениям программное ядро.

Мы разрабатываем анализатор для языков алтайского типа; описанный метод не является оптимальным для агглютинативных языков.

Особенности агглютинативных языков:

– развитая система словоизменительных аффиксов, большинство из которых грамматически однозначны (т.е. одним аффиксом выражается один грамматический признак);

– единый тип словоизменения: отсутствие строгого разграничения между именным и глагольным типом словоизменения – склонением и спряжением (ср. флективные языки);

– отсутствие значимых морфонологических чередований в основах, четкая фонетическая обусловленность использования алломорфов.

Агглютинативная словоформа образуется путем присоединения к основе в строгом порядке однозначных стандартных аффиксов; границы морфем отчетливы, фонетические изменения на стыках морфем подчиняются строгим правилам. Но попытки построить парадигму конкретного слова демонстрируют ее чрезвычайную сложность и многоместность, что обусловлено большим числом словоизменительных аффиксов. Это подталкивает нас к построению морфологического анализатора, учитывающего все возможные в тюркских языках комбинации морфем. Существующие тюркские парсеры строят редуцированные парадигмы, ограничивая число словоформ примерно до трехсот, что пагубно сказывается на качестве их работы.

Для построения компьютерной модели тюркской словоформы был использован подход, традиционно применяющийся отечественными исследователями при описании агглютинативных языков, особенно в полевых условиях. В лингвистике этот подход принято называть «грамматикой порядков» (см. [2]). Грамматика порядков – удобный инструмент описания агглютинативных языков, морфология которых соответствует

следующим требованиям: а) фиксированная последовательность словообразовательных аффиксов; б) их грамматическая однозначность; в) однократность появления в данной словоформе аффикса определенной граммемы.

1. Основные компоненты морфологического анализатора

В построении морфологического анализатора задействованы три основных лингвистических компонента:

- словарь языка (словарь, содержащий частеречные пометы и чередования основ, не описанные фонологическими правилами);
- компьютерная модель словоформы, опирающаяся на адекватное грамматическое описание (ориентированное на автоматический анализ языка);
- набор правил сочетаемости, включающий правила сочетаемости аффиксов в пределах словоформы и фонетические правила выбора алломорфов конкретного аффикса.

Алгоритм анализа, разработанный Ф.Крыловым, позволяет заполнять вышеперечисленные позиции материалами любого тюркского языка. Анализ словоформы идет справа налево. Сначала программа ищет в словаре основ целую словоформу. Если ее там не оказывается, парсер ищет с правого конца словоформы словоизменительный формант и, если обнаруживается последовательность символов, похожая на аффикс из базы, она отрезается и проходит проверку на возможность следования непосредственно за основой, а левая часть снова сравнивается со словарем основ. При положительном результате парсер предлагает для такой словоформы вариант анализа. При отрицательном результате программа снова обращается к правому концу словоформы и ищет следующий формант, сравнивая его с базой аффиксов. Так продолжается до тех пор, пока оставшаяся слева часть словоформы не совпадет со словом из словаря основ. К настоящему времени работает версия анализатора для хакасского языка [3]; в разработке анализаторы для шорского, тувинского, якутского и древнетюркского языков. Словари основ для этих языков автоматически извлечены с помощью СУБД STARLING [4] из распознанных и выправленных электронных копий больших тюркско-русских словарей.

Ниже описана версия морфологического анализатора для древнетюркского языка.

2. Модель словоформы и правила автоматического анализа древнетюркского языка. Словарь основ построен на Древнетюркском словаре [5]; постепенно ведется его дополнение по словарю Клосона [6].

NB, что для древнетюркского последовательность слотов именных категорий дублирована. Это сделано из-за многих явлений: двойных падежей, возможности изменения по падежам атрибутивизированных падежных форм, возможности изменения по падежам комитатива и делибератива и некоторых других. Примеры: *mäniŋlär ol* ‘они мои’, *mäniŋsiz* ‘состояние "не-я" (букв. "без моего")’, *öziniŋčä* ‘как его собственный’, *biziŋtäkičä* ‘как в одном из наших’, *barmišlar-niŋ-ï-nda* ‘от одного из тех, кто шел’, *kutïnlïγ* 'принадлежащий к его богатству' [7].

Морфы в строке морфов и граммемы в строке граммем разделяются дефисами.

Кумулятивно выраженные граммемы разделяются точками, субморфы внутри морфа (аффикса, занимающего слот) также разделяются точками.

Части композита разделяются знаком +.

Условные обозначения

S – основа.

1. **"Глагольные" слоты**

Neg – отрицание

Tense, Mood – время, наклонение

Sequ – секвентатив, действие, предшествующее главному

Praes – презенс

Fut – будущее время (в чем семантическая. разница первого и второго будущего, неясно, возможно, они распределены по диалектам)

Indir – индиректив, прошедшее время (перфект) с косвенной эвиденциальностью)

Perf – перфект (действие в прошлом, результат которого имеется в настоящем), не маркированнный по эвиденциальности

Res – результатив (перфект, маркированный по прямой эвиденциальности – это либо действие, свидетелем которого был говорящий, либо действие, имеющее результат, наблюдаемый в настоящем)

Praet – претерит (немаркированное прошедшее)

FutIm – непосредственное будущее

Inf - инфинитив

PrtImpf – имперфективное причастие

PrtAct – активное причастие

PrtHab – хабитуальное причастие

PrtAuct – агентивное причастие

PrtProsp – проспективное причастие

PrtProj – проективное причастие

PrtNecess – причастие необходимого следствия

Conv – деепричастие (семантическая разница 1 и 2 неясна, 2 употр. гораздо реже)

ConvFin – деепричастие цели

ConvDelim – деепричастие ограниченного действия

Cond – кондиционалис, условное наклонение

Imp – императив, повелительное наклонение

2. **"Именные" слоты**

Num – число. NB: Sg, единственное число отдельно не маркируется, бывает только кумулятивно с лицом или принадлежностью.

Pl - множественное число

Poss – посессивность, принадлежность

Poss1 – 1 лицо посессора

Poss2 – 2 лицо посессора

Poss3 – 3 лицо посессора

Coord – координатив (при перечислении однородных членов предложения)

Simple declension – набор падежных аффиксов простого склонения

Possessive declension – набор падежных аффиксов притяжательного склонения (после показателя принадлежности)

Список падежей

Nom – основной падеж, не маркируется.

Gen – генетив (родительный)

Dat – датив

Acc – аккузатив (винительный)

Loc – локатив (местный)

Abl – аблатив (исходный)

Instr – творительный

Equ - экватив

Dir – директив

Part – директив-партитив

Simil - симилятив

Comit – комитатив

Atr – атрибутивизатор

Модификаторы

Rel – релативизатор

Priv – приватив

Comp – компаратив

Dimin – диминутив

Финитный слот

Pers – лицо предиката главной предикации

1, 2, 3 – лица

Emph – эмфатическая частица

2.1. Ограничения на сочетаемость аффиксов

1. Слоты (позиции) 1-2 могут заполняться только у слов с пометой Verbum; заполнение слотов 3-15 возможно для таких слов, только если заполнен слот 2.

2. Кумулятивные показатели Imp.Pers и Praet.Pers. состоят из заполнения слота 2 + заполнения слота 15; поэтому могут стоять только непосредственно за заполнением слотов 0 или 1 и только у слов с пометой Verbum.

3. Аффиксы посессивного склонения употребляются только в словоформах с заполненными слотами 4 (Poss1) или 10 (Poss2).

2.2. Правила выбора алломорфов

Гласный в скобках: проясняется, если предыдущий морф кончается на согласный, опускается, если предыдущий морф кончается на гласный.

Согласный в скобках: проясняется, если предыдущий морф кончается на гласный, опускается, если предыдущий морф кончается на согласный.

У посесс. афф. 3 л *n-* выступает обязательно перед гласной и факультативно перед согласной и # (пауза, диэрема).

Если встречаются две скобки:

-sI(n)-(n)Xŋ >-sIn-Xŋ

-sI(n)-(X)n >-sI-n

Т.е. прояснение или опущение буквы в скобках отсчитывается слева направо.

Если внутри одной клетки стоят несколько аффиксов, то это варианты аффиксов с невыясненными позициями появления, т.е. считающиеся свободными вариантами.

Таблица 1

Модель древнетюркской словоформы и набор древнетюркских словоизменительных аффиксов

	0	1	2	3	4	5	6		7	8	9	10	11	12		13	14	15
№							Case$_1$							Case$_2$				
№ п/п	S	Neg	Tense/ Mood	Num$_1$	Poss$_1$	Apos$_1$	Simple declension	Possessive declension	Atr$_1$	Comit$_1$	Num$_2$	Poss$_2$	Apos$_2$	Simple declension	Possessive declension	Atr$_2$	Comit$_2$	Person
1.		Neg *-mA*	Neg.Indir *–mA.dUk*	Pl - *lAr*	Poss1.sg *-(X)m*	Apos *-lI*	Gen - *(n)Xŋ* Gen *-nXŋ* Gen *–nXg* Gen *-nUŋ*	Gen - *(n)Xŋ*	Atr *-kI*	Com *-lXg*	Pl - *lAr*	Poss1.sg *-(X)m*	Apos *-lI*	Gen - *(n)Xŋ* Gen *-nXŋ* Gen *–nXg* Gen *-nUŋ*	Gen - *(n)Xŋ* *ıo*	Atr *-kI*	Com *-lXg*	1.sg – *mAn* 1.sg – *bAn*
2.			Neg.Sequ *–mAtIn* Neg.Sequ *–mAtI*		Poss2.sg *-(X)ŋ* Poss2.sg *-(X)g*		Acc *–(X)g* Acc *–nI*	Acc *–nI* Acc *–In*		Delib *-sXz*		Poss2.sg *-(X)ŋ* Poss2.sg *-(X)g*		Acc *–(X)g* Acc *–nI*	Acc *–nI* Acc *–In*		Delib *-sXz*	2.sg *-sAn*
3.			Neg.Praes *–mAz* Neg.Praes		Poss3 - *(s)I(n-)*		Dat *–kA* Dat *-gA*	Dat *–kA* Dat *-gA* Dat *–A*		Comp *-dAg*		Poss3 - *(s)I(n-)*		Dat *–kA* Dat *-gA*	Dat *–kA* Dat *-gA* Dat *–A*		Comp *-dAg*	1.pl *–bXz* 1.pl – *mXz*

			–mAs														
4.			Neg.Fut$_2$ *–mAčI*		Poss1.pl *-(X)mXz* Poss1.pl *-(U)mUz*		Loc *–tA* Loc *-dA*	Loc *–tA* Loc *–dA*		Di-min *–kyA* Di-min *-kIñA*		Poss1.pl *-(X)mXz* Poss1.pl *-(U)mUz*		Loc *–tA* Loc *-dA*	Loc *–tA* Loc *–dA*		Di-min *–kyA* Di-min *-kIñA*	2.pl *–sXz* 2.pl.Pol *-sXz.lAr*
5.			Indir *–mIš*		Poss2pl *-(X)ŋXz* Poss2pl *-(X)gXz*		Abl *–dIn* Abl *–tIn* Abl *–dAn* Abl *–tAn*	Abl *–dIn* Abl *–tIn* Abl *–dAn* Abl *–tAn*				Poss2.pl *-(X)ŋXz* Poss2.pl *-(X)gXz*		Abl *–dIn* Abl *–tIn* Abl *–dAn* Abl *–tAn*	Abl *–dIn* Abl *–tIn* Abl *–dAn* Abl *–tAn*			3.Pl *-lAr*
6.			Sequ *–(X)p* Sequ *–(X)p.An*		Poss2. pl.Pol *-(X)ŋXz. lAr* Poss2. pl.Pol *-(X)gXz. lAr*		Instr *–(X)n*	Instr *–(X)n*				Poss2. pl.Pol *-(X)ŋXz. lAr* Poss2. pl.Pol *-(X)gXz. lAr*		Instr *–(X)n*	Instr *–(X)n*			Imp.1.Sg *–(A)yIn*
7.			Perf *–dUk* Perf *–tUk*				Equ *-čA*	Equ *–čA*						Equ *-čA*	Equ *–čA*			Imp.2.Sg *–0* Imp.2.Sg. Emph$_1$ *–gIl* Imp.2.Sg. Emph$_2$ *–čU*
8.			Res *–yUk*				Dir *–gArU*	Dir *–gArU* Dir *–ArU*						Dir *–gArU*	Dir *–gArU* Dir *–ArU*			Imp.3 *–zUn* Imp.3 *–sUn* Imp.3 *–čUn*
9.			[Praet *–d* Praet *–t*]				Part *-rA*	Part *–rA*						Part *-rA*	Part *–rA*			Imp.1.Pl *–(A)lIm*
10.			Praes*–(y)Ur* Praes *–(I)r* Praes *–Ar*				Simil *-lAjU* Simil *–čU.lAjU*	Simil *-lAjU* Simil *–čU.lAjU*						Simil *-lAjU* Simil *–čU.lAjU*	Simil *-lAjU* Simil *–čU.lAjU*			Imp.2.Pl *-(X)ŋ*
11.			Fut$_1$ *–gAy* Fut$_1$ *–kAy*				Comit *–lXgU*	Comit *–lXgU*						Comit *-lXgU*	Comit *–lXgU*			Imp.2. Pl.Pol *-(X)ŋ.lAr*
12.			Fut$_2$ *–dAčI* Fut$_2$ *-tAčI*															Imp.3.Pl *–zUn.lAr* Imp.3.Pl *–sUn.lAr* Imp.3.Pl *-čUn.lAr*
13.			FutIm *–gAlIr* FutIm *-kAlIr*															Praet.1.sg *–d.Xm* Praet.1.sg *–t.Xm*
14.			Inf *-mAk*															Praet.2.sg *–d.Xŋ* Praet.2.sg *–d.(X)g* Praet.2.sg *–t.Xŋ* Praet.2.sg *–t.(X)g*
15.			PrtImpf *–(X)gmA*															Praet.3 *–d.I* Praet.3 *–t.I*
16.			PrtAct *-(X)glI*															Praet.1.pl *–d.XmXz* Praet.1.pl *–d.UmUz*

																		Praet.1.pl *–t.XmXz* Praet.1.pl *–t.UmUz* Praet.1.pl *–d.Xk* Praet.1.pl *–t.Xk*
17.			PrtHab - *gAn* PrtHab - *kAn*															Praet.2.pl *–d.XŋXz* Praet.2.pl *–d.XgXz* Praet.2.pl *–t.XŋXz* Praet.2.pl *–t.XgXz*
18.			PrtAuct – *gUčI*															Praet.2. pl.Pol – *d.XŋXz. lAr* Praet.2. pl.Pol – *d.XgXz. lAr* Praet.2. pl.Pol – *t.XŋXz. lAr* Praet.2. pl.Pol – *t.XgXz. lAr*
19.			PrtProsp – *sXk*															Praet.3.pl *–d.I.lAr* Praet.3.pl *–t.I.lAr*
20.			PrtProj – *gU* PrtProj - *kU*															
21.			PrtNecess *–gU.lXk* PrtNecess *–kU.lXk*															
22.			$Conv_1$ – *(y)U* Conv *–A* Conv *–I*															
23.			$Conv_2$ - *(X)yIn*															
24.			ConvFin - *gAlI*															
25.			ConvDe-lim -*gInčA*															
26.			Cond *–sAr* Cond -*sA*															

2.3. Правила сандхи

2.3.1. Морфонологические процессы, проходящие на границах морфем

Сингармонизм

Гласные морфонемы:

X= i,ï, ü,ü *I= i, ï*

U = ü,u, *A=ä, a*

Беглые гласные .

В **двусложных** нечленимых основах со вторым узким гласным, отмеченных в словаре как чередующиеся, выбирается редуцированный вид основы при присоединении аффиксов, начинающихся на гласный.

2.3.2. Фонетические процессы, происходящие в словоформе

-n + ga =-ŋa

-ŋ + ga= -ŋa

* * * * * * * * * *

1. Vorislamische Alttürkische Texte: Elektronisches Corpus // URL: http://vatec2.fkidg1.uni-frankfurt.de/ (дата обращения 23.11.2013).
2. *Gleason H.* Introduction to descriptive linguistics. – New York: Holt, Rinehart and Winston, 1955. – P. 503.
3. Анализатор для хакасского языка // URL: http://khakas.altaica.ru (дата обращения 23.11.2013).
4. СУБД STARLING // URL: http://starling.rinet.ru/program (дата обращения 23.11.2013).
5. Древнетюркский словарь. / А.Боровков, В.Неделяев, Д.Насилов, Э.Тенишев, А.Щербак. – Л.: «Наука», Ленинградское отделение, 1969. – 677с.
6. *Clauson G.* An Etymological Dictionary of Pre-Thirteenth-Century Turkish. – Oxford: Clarendon Press, 1972. – P. 1022.
7. *Erdal M.* A Grammar of Old Turkic. – Leiden: Brill, 2004. – P. 576.

AUTOMATIC MORPHOLOGICAL ANALYSIS FOR CORPORA OF TURKIC LANGUAGES

A.V.Dybo, A.V.Sheymovich

This paper describes the main principles on which the automatic morphological analyzer for Turkic languages operates. Its main components are: a grammatical dictionary; a range model of a word form (including a set of ranges with a series of morphophonological forms of inflectional affixes for each range); a set of compatibility rules for affixes and a two-level set of phonetic rules that constrain the choice of components of a word form. The algorithm of automatic morphological annotation was developed by Ph.Krylov by using StarLing database processing system. The automatic morphological analyzer for the Old-Turkic language is shown below as an example. This research follows the framework of the RAS corporate project with regards to the development of corpora for languages of the Russian Federation, including Turkic minority languages.

Key words: corpus of a language, morphological analysis, morphological parser, inflection, computational model of a word form, StarLing system.

* * * * * * * * * *

1. Vorislamische Alttürkische Texte: Elektronisches Corpus // URL: http://vatec2.fkidg1.uni-frankfurt.de/ (data obrashhenija 23.11.2013). (In German)
2. *Gleason H.* Introduction to descriptive linguistics. – New York: Holt, Rinehart and Winston, 1955. – P. 503.
3. Analizator dlja hakasskogo jazyka // URL: http://khakas.altaica.ru (data obrashhenija 23.11.2013). (In Russian)
4. SUBD STARLING // URL: http://starling.rinet.ru/program (data obrashhenija 23.11.2013).
5. Drevnetjurkskij slovar' / A.Borovkov, V.Nedeljaev, D.Nasilov, Je.Tenishev, A.Shherbak. – L.: «Nauka», Leningradskoe otdelenie, 1969. – 677s. (In Russian)
6. *Clauson G.* An Etymological Dictionary of Pre-Thirteenth-Century Turkish. – Oxford: Clarendon Press, 1972. – P. 1022.
7. *Erdal M. A* Grammar of Old Turkic. – Leiden: Brill, 2004. – P. 576.

* * * * * * * * * *

Дыбо Анна Владимировна – доктор филологических наук, чл.-корр. РАН, зав. Отделом урало-алтайских языков Института языкознания РАН.

125009, Москва, Б.Кисловский пер. 1.
E-mail: adybo@mail.ru

Dybo Anna Vladimirovna – Doctor of Philology, RAS corresponding member; The Institute of Linguistics, Russian Academy of Sciences, Chief of the Department of Ural-Altaic languages.
1B.Kislovsky Per, Moscow, 125009. .
E-mail: adybo@mail.ru

Шеймович Александра Валерьевна – младший научный сотрудник Отдела урало-алтайских языков Института языкознания РАН.

125009, Россия, Москва, Б.Кисловский пер. 1.
E-mail: asheimovich@yandex.ru

Sheymovich Alexandra – The Institute of Linguistics, Russian Academy of Sciences, junior research assistant of the Department of Ural-Altaic languages.

1B.Kislovsky Per, Moscow, 125009.
E-mail: asheimovich@yandex.ru

Поступила в редакцию 12.03.2014