Two-Level Qazan Tatar Morphology

Ercan Gökgöz
Computer Engineering Department,
Fatih University, Turkey
ercangokgoz@gmail.com

Atakan Kurt
Computer Engineering Department,
Fatih University
akurt@fatih.edu.tr

Kalmamat Kulamshaev
Contempoary Turkish Dialects & Lieteratures Dept
Fatih University, Turkey
kkulamshaev@fatih.edu.tr

Mehmet Kara
Contemporary Turkish Dialects and Literatures Department
Istanbul University, Turkey
mehkara@yahoo.com

Abstract: In this paper we present a two level description of Tatar Language. Tatar is a Turkic language and the official language of Tataristan. It is spoken by millions of people mostly in the world. We describe the Tatar orthography using two level rules of Koskenniemi. These orthographic rules governing the phonology of the language during word formation is essential to morphological parsing and generation. We then represent the Tatar morphotactics using finite state machines. The FSMs for nominal and verbal morphotactics describe in detail how the words of the language can be formed. The orthographic rules and morphotactics are implemented in the Dilmac Machine Translation Framework by encoding them in XML files in an language independent way.

Key Words: Qazan Tatar morphology, orthographic rules, two-level morphology, finite state machines.

Introduction

Turkic languages are spoken by more than 200 million people in a vast geographic area stretching from Eastern Europe to China. Azerbaijani, Kazakh, Turkmen, Kyrghyz, Uzbek, Tatar, Uygur dialects are among the most spoken languages after Turkish. All Turkic languages except Turkish are computationally resource poor languages. Computational linguistics studies on these languages are very scarce. Turkish morphology was studied by Oflazer [101]. Turkmen morphology by Maxim et al. [90], and Tantug [91]. Azerbaijani by İlyas [92].

Tatar belongs the Idel-Ural (Volga-Urals) region of Kipchak subgroup of Turkic Languages [93]. Tatar, more specifically *Tatar Turkish* or *Qazan Tatar*, is the official language of the Republic of Tatarstan in Russian Federation. Tatar is spoken by more than 5 million people in Russia. There are about 10 million Tatars in Central Asia, parts of Europe and Turkey. Today Tatar language has 3 dialects: Western, Eastern and Middle. The middle dialect is spoken by Qazan Tatars. Tatars had used Arabic script until first quarter of 20th century. Current Tatar alphabet is based on the Cyrillic alphabet with some additional letters.

---

[90] M. Shylov, "Dilmaç: Turkish and Turkmen Morphological Analyzer and Machine Translation Program," Master's thesis, Fatih University, İstanbul Turkey, 2008.

[91] Tantuğ, A. C., Adalı, E., and Oflazer, K. 2006. **Computer analysis of the Turkmen language morphology.** Advances in natural language processing, proceedings (Lecture notes in artificial intelligence), 4139 . pp. 186-193.

[92] Hamzaoglu, İ. 1993. Machine translation from Turkish to other Turkic languages and an implementation for the Azeri language. MSc Thesis, Bogazici University, Istanbul

[93] Oner, M., 2007, (In Turkish) Tatar Turkcesi; Turk Lehceleri Grameri Ed., Ahmet Ercilasun, Akcag Publications, Ankara, Turkey.

Turkic languages are agglutinative languages where many inflectional and derivational morphemes are attached to root to express syntactic and semantic information. These morphemes allow one to create potentially infinite number of words [94].

Tatar like other Turkic Languages is a resource poor language. Studies on Tatar morphology are virtually non-existent. Books and articles on this language is usually in Tatar or Russian, and not available in English [95]. In this study we aim to describe Tatar morphology from the computational linguistics perspective using two-level model. The paper is organized as follows: In Section 2 the Tatar phonology will be described using two level orthographic rules. The orthographic rules describe the phonetic changes occurring when affixing morphemes to words. In Section 3 Tatar morphotactics will be described from computational point of view using Fine State Machines. In Section 4 conclusion and future work will be discussed.

Orthographic Rules of Tatar

Orthography specifies standardized path of writing system of the language. Orthography is produced by standardized orthographic rules, although sometimes includes ambiguities. These ambiguity is usually occurs in loanwords.

These two level rules are describes phonologic events during word formation when morphemes are affixed to a stem or a root. The two levels are lexical and surface level of a word. Lexical level is a formulation of a morphological parsing of a word in a written text. In lexical level the root word and the sequence of morphemes affixed to are represented such as Noun + Plural + 1PersonPossesive. The surface level of a word is the word as it appears in the text. Parsing is the process of attaining of lexical level from the surface level of a word. The rules and meaning are given in **Table 1**.

Table 1 Orthographic Rules

Syntax	Meaning
a:b ⇨ lc_rc	Lexical a is realized as surface b, only when converion's left side equals to lc and the right side equals to rc
a:b ⇦ lc_rc	Lexical a is always realized as surface b, when converion's left side equals to lc and the right side equals to rc
a:b ⇔ lc_rc	Lexical a always and only realized as surface b, when converion's left side equals to lc and the right side equals to rc
a:b /⇦ lc_rc	Lexical a is never realized as surface b when converion's left side equals to lc and the right side equals to rc

Tatar Alphabet

Tatar is written in Cyrillic alphabet. It is also written in unofficial Latin. In the past Tatars used Arabic script until the revolution in 1917. In this study we will use the following Latin Tatar alphabet consisting of 35 letters which 9 of is vowel given in **Table 2.**

Vowels are a,e,ı,i,o,ö,u,ü,é. Consonants are b,v,g,d,n,j,z,h,y,k,l,m,y,u,y,a,p,r,s,t,u,f,x, ç,ş,ç,ş,c,ñ.

Table 2 Tatar Alphabet

Cyril	Latin	Cyril	Latin	Cyril	Latin
А а	A a	Ү ү	Ü ü	Ф ф	F f
Ә ә	E e	Л л	L l	Һ	H h
Б б	B b	М м	M m	Х х	X x
В в	V v	Н н	N n	Ц ц	Ts ts
Г г	G g	Ң	Ñ	Ч ч	Ç ç
К к	K k	О о	O o	Ш ш	Ş ş

[94] Tatar Turkcesi; Prof. Dr. Mustafa Oner, Turk Lehceleri Grameri, Prof. Dr. Ahmet Ercilasun, Akcag, 2007.

[95] Poppe, N. N. (1963). Tatar manual: descriptive grammar and texts with a Tatar-English glossary. Bloomington: Indiana University.

Д д	D d	Ө ө	Ö ö	Щ щ	Şç şç
Е е	Éé,yé	П п	Р р	Ы ы	I ı
Ё ё	Yo yo	Р р	R r	І і	İ i
Ж ж	J j	С с	S s	Э э	Е е
З з	Z z	Т т	T t	Ю ю	Yu yu
И и	İy iy,İi	У у	Uw uw	Я я	Ya ya
Й	Y y				

Tatar employs vowel harmony like other Turkic languages. Like other Turkic Languages Tatar has consonant softening, consolidation and harmony, assimilation, vowel conversion, vowel drop, vowel epenthesis, consonant duplication. Below are lexical meta morphemes used in two level rules:

Consonants : C =(y,b,k,f,v,l,h,g,m,d,n,ç,ş,j,p,c,z,r,s,h,t)
Vowels: V = (a,e,é,ı,i,o,ö,ü,u)
Front Vowels: Vf = (e,i,é,ü,ö)
Back Vowels: Vb = (a,ı,o,u)
A = (a,e)
H = (ı,é)
I = (ı,i)
U = (ü,u)
L = (l,d)
M = (m,n,ñ)
P = (p,b)
G = (k,g)
D = (d,t)

1. a : ı ⇨__ +:0 y

The lexical a at the end of a word is converted to ı if the preceding affix starts with y.

Lexical: sayra+y V(caw) VVI_TAORSH
Surface: sayrı0y sayrıy (to be caw)(ötmek)

Lexical: sırla+ym V(draw) VVI_TAORSH
Surface: sırlı0ym sırlıym (draw cavity lines)(oyuk çizgiler çizmek)

4. L:n ⇨ M+:0__Ar

The lexical L is converted to n, if the word ends with m, ñ or n, and the preceding affix is LAr

Lexical: ülen+LAr N(grass)+NNI_PUL
Surface: ülen0ner ülenner (grasses)(otlar)

Lexical: urman+LAr N(forest)+ NNI_PUL
Surface: urman0nar urmanlar (forests)(ormanlar)

8. p:b ⇨ __+:0V

The lexical p at the end of a morpheme is converted to b if the preceding affix starts with a vowel.

Lexical: üp+er N(kiss)+ VVI_TAORSH
Surface: üb0er öper(kisses)

Lexical: küp+rAk N(more)+NNI_POSS3S
Surface: küb0érek kübérek(more than)(daha çok)

9. D:t ⇨ [f|s|t|k|ç|ş|h|p]+:0__

If a word ending with f, s, t, k, ç, ş, h, or p is affixed with morpheme starting with D, then D is realized as t.

Lexical: yeş+DAş N(age)+ NND_DAS

Surface: yeş0teş yeşteş(contemporary)(yaşıt)

Lexical: cinayet+DAş N(murder)+ NND_DAS
Surface: cinayet0teş cinayetteş(accomplice)

11 A:a ⇨ C*V_bC*+:0C*_C*

The lexical A is converted to a if the preceding vowel is a back vowel to employ vowel harmony.

Lexical: suw+LAr N(water)+ NNI_PLU
Surface: suw0lar suwlar (waters)

Lexical: kitap+DA N(book)+ NNI_LOC
Surface: kitap0ta kitapta (in the book)

20. z:s ⇨ __+:0 s

The lexical z at the end of a word is converted to s if the preceding affix's first letter is s.

Lexical: toz+sız N(salt)+ JND_SIZ
Surface: tos0sız tossız(without salt)(tuzsuz)

Lexical: küz+séz N(eye)+ JND_SIZ
Surface: küs0séz küsséz(without eye)

21. r:0 ⇨ __+:0g

The lexical r at the end of a word is dropped, if the preceding affix starts with g.

Lexical: kitirir+ge N(bring)+ NVD_GA
Surface: kitiri00ge kitirige(to bring)(götürmeye)

Lexical: éçérér+ge N(bring)+ NVD_GA
Surface: éçéré00ge éçérége(to bring)(içirmeye)

Tatar Mophotactics

Two-level morphology [96] have been applied to many languages. Tools to implement two-level morphology such as PC-KIMMO [97] is publicly available. It was originally applied to describe finite state Finnish morphology by Koskenniemi. A detailed description with an application to English is given by Antwort [98]. Two-level or finite state model later was applied to many languages such as Japanese [99], Korean [100], Turkish [101], Arabic [102], Mongolian [103]. All these languages except Arabic are related linguistically. They are Altaic languages. Like Ural languages of Finnish and Hungarian they are agglutinative. To our knowledge, Qazan Tatar morphology is not defined before. There is a work on Crimean Tatar [104].

---

96 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 Helsinky.
97 Karttunen L, 1983, PC-KIMMO: A General Morphological Processor. In Texas Linguistics Forum 22, pp.165-186.
98 Antworth, E.L., 1990, PC-KIMMO: A Two-level Processor of Morphological Analysis, Summer Institute of Linguistics, Dallas, TX.
99 Alam, Y.S., 1983, Two-level Morphological Analysis of Japanese, Texas Linguistics Forum 22, pp. 229-252.
100 Kim, D. B., Lee S. J., Choi, K.S., and Kim, G.C., 1994. A two-level morphological analysis of Korean. In **Proceedings of the 15th conference on Computational linguistics - Volume 1** (COLING '94), pp. 535-539.
101 Oflazer, K. 1994, Two-level description of Turkish morphology, Literary and Linguistic Computing, Literary and Linguistic Computing Volume9, Issue2 pp. 137-148.
102 Arabic Finite State Morphological Analysis and Generation, In COLING-96, Cophenagen, pp. 89-94.
103 Jaimai, P., Zundui, T., Chagnaa, A., and Ock, C.Y., PC-KIMMO-based Description of Mongolian Morphology, International Journal of Information Processing Systems Vol.1, No.1, 2005 pp. 41-48.
104 Kemal Altıntas, 2000. Turkish to Crimean Tatar Machine Translation System. MSc Thesis, Bilkent University, Ankara

We describe Tatar morphology using finite state machines (FSM). A finite state machine, which in principal is a directed graph, consists of a set of states and a set of transitions among these states. Transitions are the edges of graph labeled with inflectional or derivational morphemes defining in what order those morphemes can be affixed to a word. The immediate states represent words and their part of speech tagging. The initial states represent the roots words from a lexicon and their part of speech such as noun, verb, adverb, adjective, etc. The final states represent words that cannot take any ore morphemes. We define the nominal, verbal and adverbial morphotactics of the language using this FSM model. In Figure 1 only a small portion of FSM is shown because of space limitation.

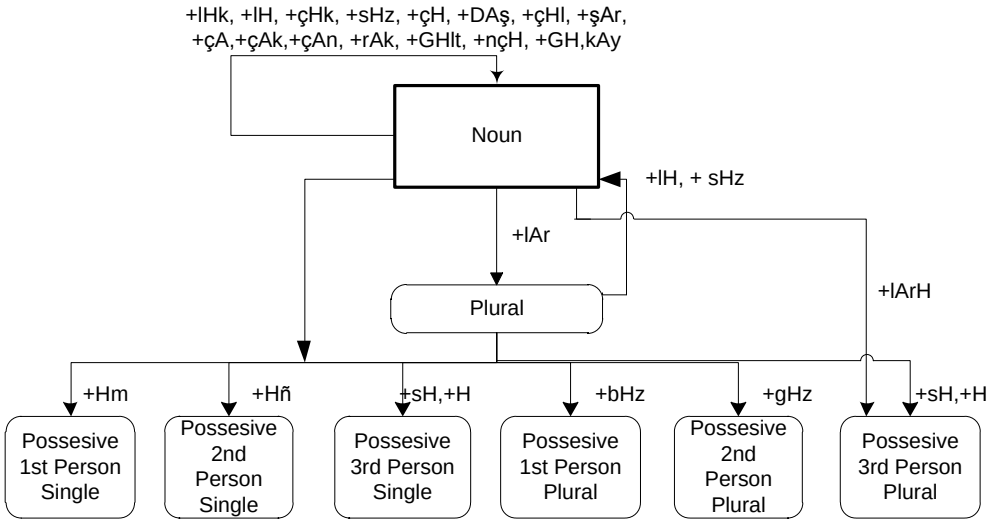

Figure 1 Nominal Morphotactis (Partially given)

Conclusions

We described Tatar morphology using the two-level morphology model and finite state machines. A number of two level orthographic rules are created to handle the conversion from surface to lexical level of a word during morphological parsing. Finite state machines for representing nominal and verbal morphotactics are given for Tatar. The model is being implemented in Dilmaç machine translation system . We conducted extensive testing of nominal and verbal Tatar conjugations. Our final objective is to implement a morphologic machine translation system between Tatar and Turkish.
105

[58] Antworth, E.L., 1990, PC-KIMMO: A Two-level Processor of Morphological Analysis, Summer Instıtute of Linguistics, Dallas, TX.
[59] Alam, Y.S., 1983, Two-level Morphological Analysis of Japanese, Texas Linguistics Forum 22, pp. 229-252.
[60] Kim, D. B., Lee S. J., Choi, K.S., and Kim, G.C., 1994. A two-level morphological analysis of Korean. In **Proceedings of the 15th conference on Computational linguistics - Volume 1** (COLING '94), pp. 535-539.
[61] Oflazer, K. 1994, Two-level description of Turkish morphology, Literary and Linguistic Computing, Literary and Linguistic Computing Volume9, Issue2 pp. 137-148.
[62] Arabic Finite State Morphological Analysis and Generation, In COLING-96, Cophenagen, pp. 89-94.
[63] Jaimai, P., Zundui, T., Chagnaa, A., and Ock, C.Y., PC-KIMMO-based Description of Mongolian Morphology, International Journal of Information Processing Systems Vol.1, No.1, 2005 pp. 41-48.
[64] Kemal Altıntas, 2000. Turkish to Crimean Tatar Machine Translation System. MSc Thesis, Bilkent University, Ankara