A Morphological Analyser for Crimean Tatar

Kemal Altintas **Ilyas Cicekli**
kemal@cs.bilkent.edu.tr ilyas@cs.bilkent.edu.tr
Bilkent University, Department of Computer Engineering
Bilkent, Ankara 06533 Turkey

Abstract : This paper describes the details of a morphological analyser designed for Crimean Tatar language. The system is based on two-level morphology and implemented using XEROX finite state tools. The phonological rules that govern the differentiation of sounds are explained in detailed. Then the morphotactic rules that organise the Crimean Tatar morpheme orders are given. A brief comparison of Turkish and Crimean Tatar is followed by some examples of the program outputs.

I. Introduction:

The language spoken by Crimean Tatars is actually a passage between Oghuz oriented Anatolian Turkish and Kipchak oriented languages such as Kazan Tatar and Kazakh-Kirgiz. Having close historical relations with Ottoman Empire, Crimean Tatar people speak a language that is intelligible to Anatolian Turks. However, the Kipchak grammar rules and words are not negligible [3].

There are three main dialects of Crimean Tatar. Northern dialect, which is called “çöl şivesi” (steppe dialect) in Crimean Tatar, shows much more Kipchak properties and is close to Kazakh and Kirgiz. The central dialect is called “Bahçesaray şivesi” referencing Bahçesaray, the capital city of Crimean Khanate and is the basic literary dialect. The southern dialect is “Yalıboyu şivesi” (coastal dialect) and is very close to Anatolian Turkish [10].

In this project, we implemented the system compatible with Bahçesaray dialect since it is the literary language. Throughout the paper, the term Crimean Tatar means “Bahçesaray dialect of Crimean Tatar language” and the term Turkish means “literary Turkish language spoken in Turkey”. Most of the root words in Crimean Tatar are common with Turkish [3, 9].

However, today, the differences both in roots and in grammatical rules are not negligible. Many words, especially in Northern dialect, are completely different from Anatolian Turkish.

Azbar : avlu (yard)
Kökrek : göğüs (chest)
Yengil : hafif (light)

Many words are present in both languages, however they mean different:

“Taşlamaq” in Crimean Tatar means to leave something at somewhere, however in Turkish “taşlamak” is to stone.
“Salmaq” in Crimean Tatar is to put or add and “salmak” in Turkish is to let something go.

There are many variances between the grammars of Turkish and Crimean Tatar. For example, the second tense of a verb is written as a separate word in Crimean Tatar while it is joined to the root in Turkish. Also, the narrative suffix in Crimean Tatar is –gen or its equivalents according to harmony rules, while narration is expressed with –miş or with its equivalence class in Turkish. For example “kelgen edi” is written as “gelmişti” (he had come) in Turkish.

Living under Russian rule for more than two centuries, the effects of Russian is heavily felt over Crimean Tatars. Not only there are many words derived from Russian, sometimes even Russian grammatical rules are applied to Crimean Tatar words. However, since these are not valid structures for Crimean Tatar, they have not been considered for the system we developed. Words, especially related to technology and usually the counterparts of Turkish words that come from western languages, are mostly derived from Russian. Some examples are:

Televizor : televizyon (television)
Avtobus : otobüs (bus)
Peçqa (peçka): soba (stove)

The rest of the paper is organized as follows: the next section gives a brief explanation of two-level morphology. The following section gives the Crimean Tatar alphabet and the fourth section lists the vowel and consonant harmony rules. The fifth section explains the morphotactics for Crimean Tatar and the sixth section compares Turkish and Crimean Tatar grammars. The seventh section summarises the implementation of the program with several example runs and the paper ends with a conclusion section.

II. Overview of Two-Level Morphology :

Two-level morphology is a way of handling morphological structures by executing pseudo-parallel rules.[2] There are two levels of the system, surface level and lexical level. Surface level representation is the direct representation of an input, as it is represented in the original language. Lexical level is the decomposed form of the input and is the output of the system when the surface representation is given as an input. In a finite state transducer, normally the surface and lexical levels are represented as two expressions separated by a colon. For example an expression like *a:b* is usually expected to mean "lexical form a is derived from the surface form b".

Rules that denote the morphological modifications and variations are all executed in parallel and all the rules work on the same input. If all of the rules accept the input, then the machine accepts the input. However, if the input is rejected by any of the rules, then the machine rejects the input directly.

There are four different rule types in such a system:

a:b => LC _ RC : Lexical *a* is mapped to surface *b* if it appears in these left and right contexts. However, its appearing in this context does not require such a mapping. In other words, if *a* is mapped to *b*, then it must be in this context and cannot happen in another context.

a:b <= LC _ RC : lexical form *a* is mapped to surface form *b* if it appears in LC and RC. However, it is also possible to map *a* to *b* in another context.

a:b <=> LC _ RC : a lexical *a* is always mapped to a surface *b* in this context and this is possible only in this context.

a:b /<= LC _ RC : a lexical *a* is never mapped to a surface *b* in the given context.

The morphotactic rules are compiled to a finite state transducer and are joined with these rules. The system as a whole tries to locate the roots and possible following suffixes for a given surface form input. If the system at any stage cannot locate a valid suffix or it discovers a situation violating the morphological modification rules, it returns with no answer. For a detailed explanation of two-level morphology, see [8].

III. The Alphabet :

As it is the case for all Turkic languages, Crimean Tatar was also written using the Arabic Script in the beginning of twentieth century. After the formation of Soviet rule, the Cyrillic alphabet of Russian language was started to be used. Now, a Latin based alphabet which is the same as Turkish alphabet with few additions was accepted by the Crimean Tatar National Assembly and is being used. The Latin based Crimean Tatar alphabet is:

Aa Ââ Bb Cc Çç Dd Ee Ff Gg Ğğ Hh Iı İi Jj Kk Ll Mm Nn Ññ Oo Öö Pp Qq Rr Ss Şş Tt Uu Üü Vv Yy Zz

In the program, the letters that are present in the ASCII characters are used as is in lowercase. Both in surface form and lexical form, we represented the letters which are absent in ASCII, with the capital form of the closest symbol. The correspondences are as follows :

ç – C ğ – G ı – I ö – O ş – S ü – U ñ – N

At the lexical level, however, we need to use some extra characters to represent one to many mappings and exceptions. For this purpose, we use the following capital letters which are used only in the program and are invisible to the user:

J – ç that does not change to c : kUJ + U = kUCU
P – p that does not change to b : saP + I = sapI
Q – q that does not change to ğ : baQ+a = baqa
T – t that does not change to d : beT + i = beti
W – k that does not change to g : teW + i = teki
H – corresponds to symbols I, i, u, U according to vowel harmony
A – corresponds to a or e according to vowel harmony rules
Y – corresponds to u or U according to vowel harmony rules.
K – corresponds to g, k, G, q according to consonant harmony rules
D – corresponds to a d or t according to consonant harmony rules.
Z – s that does not drop as a joining sound : alim + Ziñ = alimsin

We also use the following groupings in the two level morphology rules.

The vowels are	(VOWEL)	= a e I i o O u U A H M Y â
The consonants are	(CONS)	= b c C d f g G h j k l m n N p q r s S t v y z K Z B P Q J W

The other groupings are as follows :

Back Vowel	(BACKV)	= a I u o â;
Front Vowel	(FRONTV)	= e i O U;
Front Unrounded Vowel	(FRUNROV)	= i e;
Front Rounded Vowel	(FRROV)	= O U;

Back Rounded Vowel	(BKROV)	= u o;
Back Unrounded Vowel	(BKUNROV)	= a I â;
Soft Consonants	(SEDALI)	= b c d g G j v z l m n N r y h B;
Hard Consonants	(SEDASIZ)	= p C t k q S f s Z P Q J W;
Joining Consonants	(X)	= s y;

IV. Vowel and Consonant Harmony Rules :

"A realized as a"

A:a => [:BACKV] [CONS]* (%+:0) [CONS: | :CONS | :0]* _ ;

After a back vowel, the following A must be represented as an a.

bala + lAr -> balal**a**r (çocuklar – kids)

qoy + lAr -> qoyl**a**r (koyunlar - sheep)

"A realized as e"

A:e => [:FRONTV] [CONS]* (%+:0) [CONS: | :CONS | :0]* _ ;

Similar to the following rule, this follows the grammar rule stating that front vowels are to follow front vowels :

kOy + lAr -> kOyl**e**r (köyler – villages)

gUl + DAn -> gUld**e**n (gülden – from the rose)

"A realized as y"

A:y <=> [:VOWEL] %+:0 _;

In Crimean Tatar, present progressive tense suffix is –y and future suffix is -ycAK if the root ends in a vowel:

sora + A -> sora**y** (soruyor – s/he is asking)

qorCala + AcAK -> qorCala**y**caq (koruyacak – s/he will protect)

"H realized as u"

H:u => .#. [CONS]* [:BKROV] [CONS]* (%+:0) [CONS: | :CONS | :0]* _ ;

If there is only one syllable in the root, which means there is only one vowel before H and if it comes after a back rounded vowel (o, u), it is resolved to u:

soN+HncH -> soN**u**ncI (sonuncu – the last)

"H realized as U"

H:U => .#. [CONS]* [:FRROV] [CONS]* (%+:0) [CONS: | :CONS | :0]* _ ;

In other cases, namely H coming in the second syllable and following a front rounded vowel (U, O), it is resolved to U:

kOy + ZHz -> kOys**U**z (köysüz – without a village)

UC + HncH -> UC**U**nci (üçüncü – the third)

"H realized as i"

H:i => [:VOWEL] [CONS]* (%+:0) [CONS: | :CONS | :0]* [:FRONTV]
[CONS]* (%+:0) [CONS: | :CONS | :0]* _ ;
.#. [CONS]* [:FRONTV] [CONS]* (%+:0) [:CONS | CONS: | :0]* _;

In Crimean Tatar language, the u and ü in suffixes can appear only in the second syllable, and for the same suffix, it is written as ı or i in the third and the later syllables. Few exceptional morphemes, such as past morpheme –di and accusative morpheme –ni, are most of the time written with ı/i even if they appear in the second syllable. Here are two rules operationg in parallel. The first rule checks whether there are at least two syllables. If there are, then H is resolved to i after all front vowels, namely e, i, O, U. If there is one syllable, then the second rule runs and maps H to i after only front unrounded vowels.

kOr + DH -> kOrd**i** (gördü – s/he saw)

sUt + sHz + lHK -> sUts**U**zl**i**k (sütsüzlük – milklessnes)

"H realized as I"

H:I => [:VOWEL] [CONS]* (%+:0) [CONS: | :CONS | :0]* [:BACKV]
[CONS]* (%+:0) [CONS: | :CONS | :0]* _ ;
.#. [CONS]* [:BACKV] [CONS]* (%+:0) [:CONS | CONS: | :0]* _;

This is the corresponding rule for the previous one. It checks whether there are at least two syllables. If there are, it maps H to I if the previous vowel is a back vowel, namely a, â, ı, o, u. If there is one syllable, it maps H to I only if it comes after a back unrounded vowel :

azbar + HmHz -> azbar**I**m**I**z (bahçemiz – our garden)

qal + DH -> qald**I** (kaldı – he stayed)

"H is dropped after a vowel, before a morpheme"

H:0 <=> [:VOWEL] %+:0 _ ;

If H comes in the beginning of a morpheme and the last symbol of the previous morpheme is a vowel, then H drops :

eki + HncH -> ekinci (ikinci - second)

tile + Hr -> tiler (diler – s/he wishes)

"Y realized as u"

Y:u => [:BACKV] [CONS]* (%+:0) [CONS: | :CONS | :0]* _ ;

In some cases, morphemes are written with u/ü and never with ı/i. An example for such morphemes is -uv/-üv which makes nouns from verbs. Y is used and resolved into u if it follows something resolved into a back vowel:

toplaS + Yv -> toplaS**u**v (toplanış / toplanma – gathering / meeting)

oq + Yv -> ok**u**v (okuma – reading / education)

"Y realized as U"

Y:U => [:FRONTV] [CONS]* (%+:0) [CONS: | :CONS | :0]* _ ;

If Y comes after some symbol that is resolved into a front vowel, then it is resolved into U:

kel + Yv -> kel**U**v (geliş / gelme – coming)

"Y is dropped after a vowel, before a morpheme"

Y:0 <=> [:VOWEL] %+:0 _ ;

If Y comes in the beginning of a morpheme and after a vowel, it is dropped:

sayla + Yv -> saylav (seçim - election)

"K realized as k"

K:k => [:FRONTV] [:CONS]* [:SEDASIZ] %+:0 _ [:FRONTV] [CONS: | :CONS | :0]* ;
[:FRONTV] (%+:0) _ [.#.|[(%+:0) [CONS]]];

In Crimean Tatar language, there are two different k sounds which are also represented seperately in writing : one is represented by k and the other is by q. "k" is paired with front vowels and q is with back vowels. Also there are "sedalı" (soft) and "sedasız" (hard) consonants which affect their changes in the words. In morphemes, k softens to g, and q softens to ğ. Please note that the sound ğ is not the same as the soft g in Turkish and it is much harder a sound. All these forms are represented by the capital K symbol.

K corresponds to k when it comes in the beginning of a morpheme where the last sound in the root is a hard consonant and the last vowel is a front vowel or the root ends in a front vowel.

ket + Kan -> ket**k**en (gitmiş – s/he went)

kel +mAK -> kelme**k** (gelmek)

"K realized as q"

K:q => [:BACKV] [:CONS]* [:SEDASIZ] %+:0 _ [:BACKV] [CONS: | :CONS | :0]*;
[:BACKV] (%+:0) _ [.#.|[(%+:0) [:CONS]]];

If K is in the beginning of a morpheme and preceded by a back vowel and a hard consonant, or it follows a back vowel, it is reallized as q.

saC + Kan -> saC**q**an (ekmiş - planted)

baq +AcAK + mMz -> baqaca**q**mIz (bakacağız – we will look)

"K realized as g"

K:g => [[:FRONTV] [:CONS]* [:SEDALI] | [:FRONTV]] %+:0 _ [:FRONTV] [CONS: | :CONS | :0]*;
[:FRONTV] _ %+:0 (:0) [:FRONTV];

This rule and the following one are the pair stating the rules for softening the k and q. K corresponds to g if it is preceded by a front vowel and a soft consonant (sedalı) or if it is preceded and followed by front vowels.

kel + AcAK + Hm -> kelece**g**im (geleceğim – I will come)

piSir + KAn -> piSir**g**en (pişirmiş – s/he cooked)

"K realized as G"

K:G => [[:BACKV] [:CONS]* [:SEDALI] | [:BACKV]]%+:0 _ [:BACKV] [CONS: | :CONS | :0]*;
[:BACKV] _ %+:0 (:0) [:BACKV];

K is paired with G when it comes between two back vowels or it follows a back vowel and a soft consonant.

qal + AcAK + Hm -> qalaca**G**Im (kalacağım – I will stay)

al + KAn -> al**G**an (almış – s/he took)

"X is deleted after a consonant"

X:0 <=> [:CONS | CONS:] %+:0 _ ;

The symbols n, s and y are not written if they follow a consonant but written if they follow a vowel. For example s in the following morpheme is deleted :

ev + sH -> evi

"D realized as t"

D:t <=> [:SEDASIZ] %+:0 _ ;

D is realized as t if and only if it follows a symbol that corresponds to a hard consonent (sedasız). Otherwise it is realized as d.

ket + DH -> ket**t**i (gitti – s/he went)

kitap + DAn -> kitap**t**an (kitaptan – from the book)

"k realized as g"

k:g <=> [VOWEL] _ %+:0 (X:0) [:VOWEL];

The symbol k in the end of a word is realized as g if it is followed by a vowel in the following morpheme, possibly with a sound dropping in between:

yürek + Hm -> yüre**g**im (yüreğim – my heart)

eSek + sH -> eSe**g**i (eşeği – his/her donkey)

Note that this is not the same as K realized as g in the previous rules.

"q realized as G"

q:G <=> [VOWEL] _ %+:0 (X:0) [:VOWEL];

Symbol q is changed into a G if it succeeds a vowel and the beginning of the following morpheme is a vowel.There may possibly be a dropping joining sound such as s.

ayaq + sH -> aya**G**I (ayağı – his/her foot)

qaSIq + HmHz -> qaSI**G**ImIz (kaşığımız – our spoon)

"C realized as c"

C:c => [VOWEL] _ %+:0 (X:0) [:VOWEL];

The character C corresponds to a c if it comes between two vowels with a possible dropping sound.

aGaC + HmHz -> aGa**c**ImIz (tahtamız – our wood)

"p realized as b"

p:b <=> [VOWEL] _ %+:0 (X:0) [:VOWEL];

The symbol p is changed into a b if it is followed by a vowel.

kitap + sH -> kita**b**I (kitabı – his/her book)

Garip + Hm -> Gari**b**im (garibim – my poor)

"c realized as C"

c:C => [:SEDASIZ] %+:0 _ [:VOWEL];

The symbol c corresponds to a C after a root or a morpheme ending in a hard sound (sedasız). This is especially for the morpheme -cı / -ci which makes nouns from nouns.

qurt + cH -> qurt**C**u (kurtçu – wolf trainer)

aS + cH -> aS**C**I (aşçı – cook)

For the last few rules, it is necessary to state that there are exceptional cases. For example, for the word “sap + sH”, it becomes “sapI”, namely the symbol p does not change into b. Or for the word “tek”, again there is no change. However, for “tUp”, there is a change when the “sH” morpheme is added: “tUbU”. There is no strict rule for these kinds of words. The way we handle them is changing all the C's or p's whenever it is possible, and writing those words which do not change with a different symbol. For example, the word “tek” will internally be written as “teW” and sap as “saP”. Note that all the symbols are normally lower case symbols except for special Turkish characters. The rest of the uppercase characters are special cases handled in different situations.

V. Morphotactics :

V.I Roots :

The root words for this application are compiled from pieces of literary works. They include words from different dialects of the language. The reasons behind the fact that the total number of roots is not very high are various. First of all, our lexicon does not include many words derived from Russian and other languages. Only a very small part of proper names are included in the system and technical words are not considered. Moreover, Crimean Tatar language could not find a fertile area to develop during Soviet period, leaving us with a relatively small lexicon. We hope to improve the total area covered by the roots in time. The list of words are grouped as follows :

Nouns, Verbs, Adjectives, Adverbs, Proper Names, Simple Numbers, Pronouns, Connectives

V.II Morphotactic Rules For Crimean Tatar :

Morphotactics of a language determine the order of morphemes that appear in a word. Although Crimean Tatar is located basically in Kipchak group of Turkic languages, the morphotactic rules of Crimean Tatar mostly comply with those of Turkish. In other words, the morphemes themselves are sometimes different from those of Turkish, however the meaning they imply and the order they appear in the word are usually the same as Turkish.

In the system, the finite state machine starts from a start state and checks the possible constructs beginning with the root and possible following morphemes. Each list of roots and possible following morphemes are expressed in a lexicon file. If a root is matched, then the machine gives the appropriate output and goes to the next state. Below a sample part of the lexicon can be seen:

```
LEXICON NOUNS
abide+Noun:abide        POST-NOUN;
abla+Noun:abla          POST-NOUN;
aC+Noun:aJ              POST-NOUN;
acderha+Noun:acderha    POST-NOUN;
acet+Noun:acet          POST-NOUN;
...
LEXICON POST-NOUN
+A3Sg:                  PLURAL;
+A3Pl:+lAr              PLURAL;
LEXICON PLURAL
+Pnon:                  POSSESSIVE;
+P3Sg:+sH               POSS-3;
+P1Sg:+Hm               POSSESSIVE;
+P2Sg:+HN               POSSESSIVE;
+P1Pl:+HmHz             POSSESSIVE;
+P2Pl:+HNHz             POSSESSIVE;
```

For example, for a surface form word like “abidesi” (the statue of something/someone), we can think of the internal representation as “abide + sH” which is created with the help of vowel and consonant harmony rules. The system would first check the roots for *abide* and as it finds the word there, it outputs the lexical form “abide + Noun” and goes to the next state indicated by POST-NOUN. At this state, the possible morpheme accepted is +lAr. Otherwise, the system goes to the next state, PLURAL, with zero input (epsilon transition) giving the output +A3Sg. Now the output is “abide + Noun + A3Sg”. At the PLURAL state, the system recognises the input morpheme +sH and goes to the next state POSS-3 after giving the input +P3Sg. This continues until the system reaches the final state or a state that does not accept the input. If the input is not accepted, the output is not returned to the user.

Considering the variations in morphemes, the finite state machines offered for Turkish by Oflazer in [7] can be applied to Crimean Tatar. A basic explanation of differences between Turkish and Crimean Tatar is given in the next section. For nominal morphotactics, however, the Nominal-Verb states are not applicable. In Turkish, the historical verb *imek* was lost and this word is joined to the previous verb or noun. For example, in Turkish the sentence “Evde idim” (I was at home) is usually written and said as “Evdeydim”. However, in Crimean Tatar the equivalent of this verb, *emek*, is not lost and preserved as a separate verb. “Evde edim” is never written as “Evdeydim”. Thus, this part of the Turkish finite state machine is not applicable to Crimean Tatar.

For verbal morphotactics, again the finite state machine for Turkish prepared by Oflazer is mostly applicable. Some of the morphemes such as +yAyaz, +yAkoy are not present in Crimean Tatar, whereas some extra morphemes such as +Ayata (present progressive) are to be added. The main difference is again in the use of verb *imek/emek*. In Crimean Tatar, the compound tenses are written as separate words, thus the parts related to second tense in the Turkish finite state machine are not applicable to Crimean Tatar. Also the optative case is not present in Crimean Tatar, so that part is also to be omitted.

VI. Comparison of Crimean Tatar and Turkish Grammars:

Being two closely related Turkic languages, Crimean Tatar and Turkish have most parts in common. The word order and the duties of words in the sentence are most of the time similar. The roots are usually similar, but sometimes they may have different meanings in the two languages. For example the word “kaldırmak” means *to lift* in Turkish, whereas it means *to leave something at somewhere* in Crimean Tatar context.

The differences are usually in morphemes rather than in deeper levels of grammar. In other words, different morphemes are used to get the same meaning.

Below is a tabular comparison of the to grammars. This is not a complete analysis, but rather a comparison to give some idea about Crimean Tatar grammar. It covers main aspects of Turkic languages. Details of Crimean Tatar grammar are explained in [3, 4, 5, 6]. The explanations and examples on the left column are for Crimean Tatar and those on the right are for Turkish.

VI.I Alphabet :

Crimean Tatar used to be written in the Arabic based alphabet up to the first quarter of twentieth century. After the establishment of the Soviet rule, first a Latin based alphabet was used and then Crimean Tatars were forced to use a Cyrillic alphabet. During the Soviet period, everything was printed in Cyrillic alphabet. After the collapse of the Soviet Union, a Latin based alphabet was accepted by Crimean Tatar National Assembly. Now both Cyrillic and Latin alphabets are used. Newspapers and journals are printed in both alphabets.

The current alphabet is the same as Turkish alphabet. There are three letters that differ : â, ñ, q. The letter â is a sound that is between a and e as in lâle (tulip), kâğıt (paper). The letter ñ is for nasal n and is the counterpart of Ottoman "nûn-ı türkî". It is mostly used for the second person : kelesiñ (geliyorsun – you are coming), köyüñiz (köyünüz – your village), deñiz (sea), sınır (boundary). The last of these letters is used for Turkish k, however it is always paired with back vowels : qalmaq (kalmak – to stay), qurultay (meeting), qaysı (hangi - which).

VI.II Tenses :

All the tenses present in Turkish are also present in Crimean Tatar. The usages of the tenses are almost the same. The rules and usage for past tense are the same in both languages. In the tables below, the left column gives a brief structure for Crimean Tatar and the right column form Turkish. The Turkish and Crimean Tatar examples are corresponding to each other. The first line of each explanation gives the lexical morpheme and the second line is the corresponding surface morphemes. The third line, if present, is for necessary explanations.

Tables 1, 2 and 3 explain the formation of present progressive, narrative and future tenses in Crimean Tatar respectively and compare them with Turkish.

Crimean Tatar		Turkish	
-A -a/e/y -vowel harmony rules apply -can correspond to English simple present tense and present progressive tense	baq + A = baqa kel + A = kele sora + A = soray	-Hyor -(ı/i/u/ü)yor -the suffix -yor does not coincide with vowel harmony rules	bak + Hyor = bakıyor (s/he is looking) gel + Hyor = geliyor (s/he is coming) sor + Hyor = soruyor (s/he is asking)
-Ayata -eyata/ayata -used when some event is just about to start or is definitely continuing at the moment	qal + Ayata = qalayata ket + Ayata = keteyata	-no direct Turkish correspondence -the same meaning is given by present progressive	

Table 1:Present Progressive

Crimean Tatar		Turkish	
-Kan -gen/ken/Gan/qan -vowel and consonant harmony rules apply	kel + KAn = kelgen tOk + KAn = tOkken sora + KAn = soraGan saC + KAn = saCqan	-mHş -mış/miş/muş/müş	gel + mHş = gelmiş (s/he came) dök + mHş = dökmüş (s/he poured) sor + mHş = sormuş (s/he asked) ek + mHş = ekmiş (s/he planted)

Table 2: Narrative

Crimean Tatar		Turkish	
-AcAK -acaq/ecek/ycaq/ycek -vowel harmony rules apply	al + AcAK = alacaq kOr + AcAK = kOrecek sora + AcAK = soraycaq tile + AcAK = tileycek	-yAcAk -(y)acak/(y)ecek	al + yAcAk = alacak (s/he will take) gör + yAcAk = körecek (s/he will see) sor + yAcAk = soracak (s/he will ask) dile + yAcAk = dileyecek (s/he will wish)

Table 3: Future

VI.III Compound Tenses :

In Crimean Tatar, the second tense that comes to the root is not joined with the root, but written separately with the verb "emek". In Turkic languages, the second tense normally is past or narrative. So, the second tense comes after the root as "edi" or "eken". There is an exceptional case here for vowel-consonant harmony, which says that narrative suffix that comes after a vowel is written as –gen. However, here it is written as –ken. The person suffixes are added to the second tense, rather than the root whereas passive and causative suffixes are added to the root as explained in Table 4.

Crimean Tatar		Turkish	
-edi -for past tense as second tense	yazGan edin	-di -for past tense as second tense -it is joined to the root	yazmıştın (you had written)
-eken -for narrative as the second tense	yapacaq eken	-miş -narrative as second tense	yapacakmış (s/he would have done it)

Table 4: Compound Tenses

VI.IV Cases :

Although the meaning given by the case suffixes are the same as Turkish, the suffixes themselves and formation rules are different from Turkish. Tables 5, 6, 7 and 8 explain the formation of accusative, dative, genitive and instrumental cases respectively.

Crimean Tatar		Turkish	
-nı/ni -no corresponding –nu/nü -the sound n is the part of the morpheme and is never dropped -vowel harmony rules apply	ev + nH = evni qol + nH = qolnI baca + nH = bacanI	-yH or nH -(y)ı/(y)i/(y)u/(y)ü -(n)ı/(n)i/(n)u/(n)ü -the sounds y and n joining sounds and can be dropped if morpheme follows a root ending in consonant -the vowel harmony rules for Turkish apply	ev + yH = evi (the house [Acc]) kol + yH = kolu (the arm [Acc]) baca + yı = bacayı (the chimney [Acc])

Table 5: Accusative

Crimean Tatar		Turkish	
-KA -ge/ke/Ga/qa -vowel and consonant harmony rules apply	deñiz + KA = deñizge qoranta + KA = qorantağa kökrek + KA = kökrekke at + KA = atqa	-yA -(y)a/(y)e -if root ends in a consonant, the joining sound y is dropped	deniz + yA = denize (to the sea) aile + yA = aileye (to the family) göğüs + yA = göğüse (to the chest) at + yA = ata (to the horse)

Table 6: Dative

Crimean Tatar		Turkish	
-nHN -nıñ/niñ -no corresponding -nuñ/nüñ -the sound n is the part of the morpheme and is never dropped	ev + nHN = evniN horaz + nHN = horaznIN quyu + nHN = quyunIN	-nHn -(n)ın/(n)in/(n)un/(n)ün -the sound n is a joining sound and can be dropped if morpheme follows a root ending in consonant -the vowel harmony rules for Turkish apply	ev + nHn = evin (of the house) horoz + nHn = horozun (of the hen) kuyu + nHn = kuyunun (of the well)

Table 7: Genitive

Crimean Tatar		Turkish	
-nen -no -nan form is present -le/la is rarely used under Turkish influence	Amet + nen = Ametnen avtobus+nen= avtobusnen soqur + nen = soqurnen	-(y)le/(y)la -y is the joining sound and drops when the root ends in a consonant -vowel harmony rules apply	Ahmet + ylA = Ahmet'le (with Ahmet) otobüs + ylA = otobüsle (by bus) kör + ylA = körle (with the blind)

Table 8: Instrumental

VI.V Adjective Derivation :

Adjective derivation with the narrative suffix is different from Turkish both in structure and meaning. The corresponding structures are explained in Table 9.

Crimean Tatar		Turkish	
-Kan -gen/ğan/ken/qan -one meaning is "something that has already happened"	Ol + KAn = Olgen sat + Hl + KAn = satılğan bit + ken = bitken	-mHş -mış/miş/muş/müş -has the same meaning	öl + mHş = ölmüş (dead) sat + Hl + mHş = satılmış (sold) bit + mHş = bitmiş (finished)
-the second meaning is "something that is currently continuing"	çap + qan = çapqan yür + gen = yürgen	-yAn -(y)an/(y)en	koş + yAn = koşan (running) yürü + yAn = yürüyen (walking)

Table 9: Adjective Derivation

VII. Implementation :

This system is implemented using XEROX finite state tools for language engineering [11]. The two level vowel and consonant harmony rules are compiled with twolc. The lexical rules are compiled with lexc. The basic dialect of Crimean Tatar used for the lexical rules is Bahçesaray dialect. However the roots includes words from other dialects. There are a total of 5200 root words compiled from around 80.000 word text that includes pieces from different literary works and public literature.

At the moment, the program is a prototype and there is not a special interface for the ordinary user. Turkish character set is not directly used in the program and special characters are mapped as explained in the alphabet section above.

The program runs in both ways. Given the surface form, the lexical form is produced by the program. Similarly, when the lexical form is given, the corresponding surface form is produced. The mappings are not one to one due to ambiguities in the language, so it is always possible to get more than one result.

The output first starts with the root of the surface form entered by the user. Then the type of the word is given. The following morpheme for nouns is agreement morpheme and it is followed by possessive and case information. Adjectives and pronouns are similar. For verbs, the third morpheme is sense. It is followed by tense and then agreement information. Changes in the type of the word are marked with a derivational boundary (^DB) and it is followed by the new type of the word.

Below are a few examples of the system:

kelem	kel+Verb+Pos+Prog1+A1Sg (geliyorum – I am coming)
eviNiz	ev+Noun+A3Sg+P2Pl+Nom (eviniz – your house)
qalacaGIm	qal+Verb+Pos+Fut+A1Sg (kalacağım – I will stay) qal+Verb+Pos^DB+Adj+FutPart+P1Sg (kalacağım [adjective] – [of the place] related to my stay) qal+Verb+Pos^DB+Noun+FutPart+A3Sg+P1Sg+Nom (kalacağım [bana ait olan kalma eylemi] – my prospective stay) qal+Verb+Pos^DB+Noun+FutPart+A3Sg+Pnon+Nom^DB+Verb+Zero+Pres+A1Sg (kalacak olanım – I am the one who will stay) qal+Verb+Pos^DB+Noun+FutPart+A3Sg+P3Sg+Nom^DB+Verb+Zero+Pres+A1Sg (kalacağıyım – I am his prosepective stay)
yazGanlarGa	yaz+Verb+Pos^DB+Noun+PastPart+A3Pl+Pnon+Dat (yazanlara – to those who have written)

VIII. Conclusion :

Although there is much influence of Anatolian Turkish over Crimean Tatar, it is a prototype for Kipchak oriented Turkic languages. Rules and systems developed for Crimean Tatar may easily be applied to other Kipchak languages such as Kazan Tatar, Kazakh or Kirgiz. Having morphological analysers ready in hand, we expect machine translation among these languages to be relatively easy.

Preparing the rules themselves is not very difficult. However, in some cases, the grammatical rules themselves are not very clear. Having to write their language with Cyrillic letters, and after a long period of obligatory Russian education, many of the Turkic people lost the nuances in their languages. Most of the time, their use of their mother tongue was limited to daily life issues. Sometimes, the writers of the grammar books do not agree on some rules. For example, in some books, it is said that, the known past morpheme -di is not written with u/ü. According to this, we can not use DH as past morpheme, but should use some similar morpheme such as DM instead. However some other books say that it is valid to write -du / -dü. Actually, in Crimean Tatar language, there are very few cases where the symbols u, ü, o, ö appear in the syllables after second. Thus, even if we use DH, most of the time it will be represented as dI/di due to this nature of the language itself.

Also, in this project, we do not consider numbers, proper names and the words that are derived from some foreign languages especially from Russian. Since Crimean Tatars have lived under Russian rule for about two hundred and thirty years, there are many Russian words appearing in the language. However, it is a very deep subject, which requires a considerable knowledge of Russian grammar and vocabulary.

There are one or two rules about which we could not find exact explanations and did not include in our rules. For example, the sound gemination as in sIrrI, hakkI etc. is not included. We hope to find some accurate information about this during the research and will add them to the list.

To sum up, Crimean Tatar language is similar to Turkish in many aspects, although it has some variations. We tried to cover largest possible rules for a simple pure Crimean Tatar text, without any rule abiding proper names or foreign words. There is a lot of work to do in the area and during the ongoing research, we will find out the missing parts and recover them.

References:

[1] Abdullayeb, E., Umerov, M. Russko-Krimskotatarskiy Uçebniy Slovar. QırımOquv-Pedagogik Neşriyatı, Aqmescid, 1994

[2] Antworth, Evan L. PCKIMMO: A Two-level Processor for Morphological Analysis. Summer Institute of Linguistics, Dallas, Texas, 1990

[3] Asanov, Ş. A., Garkavets, A. N., Useinov, S. M. Krimskotatarsko-Russkiy Slovar. Radyanska Shkola, Kiev, 1988

[4] Çobanzade, Bekir. Qırımtatar İlm-i Sarfı. Qırım Hükümet Neşriyatı, Aqmescid, 1925

[5] Dermenci, E., Şemsedinoma, A. Qırımtatar Tili Dersligi. Qırım ASSR Devlet Neşriyatı, 1940

[6] Memetov, Ayder. Tatar Tili Grammatikasınıñ Praktikumı. Okituvçı, Taşkent, 1984

[7] Oflazer, Kemal. Two-level Description of Turkish Morphology. Literary and Linguistic Computing, Vol. 9, No:2, 1994

[8] Oflazer, Kemal. Morphological Analysis, chapter in Syntactic Wordclass Tagging Hans van Halteren, Editor, Kluwer Academic Publishers, 1998

[9] Useinov, S. M. Qırımtatarca Rusça Luğat. Dialog, Akmescid, 1994

[10] Useinov, S. M., Mireev, V. A. Izuchayte Krimskotatarskiy Yazik. Tavriya, Simferopol, 1991

[11] http://www.xrce.xerox.com/research/mltt/fst/home.html