AVRASYA Uluslararası Araştırmalar Dergisi

Cilt : 6 Sayı : 15 Sayfa: 632-648 Kasım 2018 Türkiye

Araştırma Makalesi

A SURVEY ABOUT MACHINE TRANSLATION OF TURKIC LANGUAGES

Dr. Öğr. Üye. Murat ORHUN[♣]

ABSTRACT

With developing Internet technology, it becomes important tasks to write and edit documents through the Internet supported tools. In such tasks, spell checking, auto completing words after some syllables, machine translation of words from one language to another, content summarizing and analyzing according to words and sentence structure in a file, are considered popular and important topics. With improving social communication technology and devices, analyzing continuously produced documents and translation between different languages are becoming more important day after day. It is difficult to handle such task with manually without any error and with high speed. Because of these reasons, computer supported research languages come into prominence and a new filed, which is computational linguistic, formed in science. Turkish is one of the most computational researched language in Turkic language family and a lot of projects have been done. For other Turkic languages, computational researches are still at the beginning stage. This paper describes some main properties of the Turkic languages and summarize machine translations and natural language researches that have been done so far. A unique tagging system for Turkic word is suggested to implement a machine translations system between all Turkic languages, even between other Altaic languages. The usability of the unique tags is explained with examples.

Keywords: Machine Translation, Turkic Languages, Turkish Grammar,

TÜRK DILLERININ BILGISAYARLI AKTARILMASI HAKKINDA İNCELEME

ÖZ

Internet teknolojisinin gelişmesiyle, Internet destekli ortamlarda dosya yazmak ve geliştirmek önemli işlevler olarak kabul edilmiştir. Bu çeşit işlevler, imla düzeltmek, belli bir hecelerden sonra sözcükleri otomatik biçimde tamamlamak, sözcükleri bir dilden bir başka dile bilgisayarlı aktarmak, dosyaların tümce yapısı ve kullanılan sözcüklerine göre içerik özetlemek ve analiz etmek gibi güncel ve önemli konuları kapsıyor. Sosyal iletişim teknolojileri ve araçlarının gelişmesiyle, sürekli üretilen dosyaların analiz edilmesi ve farklı diller arasında aktarılması gün geçtikçe daha çok önem kazanıyor. İnsan gücüyle bu çeşit çalışmaların hızlı ve doğru bir biçimde yapılması oldukça zordur. Bu nedenle, bilgisayar destekli dil araştırmaları önem kazanmıştır ve yeni bilimsel çalışma alanı olan bilişimsel dilbilimi şekillenmiştir. Türk dilleri ailesinde, Türkiye Türkçesi bilişimsel dilbilimi alanında en çok çalışılan dildir ve pek çok proje geliştirilmiştir. Diğer Türk dilleri ile ilgili bilişimsel dilibilimi çalışmaları henüz başlangıç düzeyindedir. Bu makalede Türk dillerinin bazı önemli özellikleri açıklanmıştır ve şimdiye kadar yapılan bilgisayarlı aktarma ve doğal dil çalışmaları özetlenmiştir. Tüm Türk dilleri, hatta diğer Altay dilleri arasında aktarma yapabilecek bir bilgisayarlı aktarma sisteminin geliştirilmesi için, tüm Türk dilleri sözcüklerinin tek çeşit etiket ile işaretlenmesi önerilmiştir ve bu çeşit etiketlerin kullanılırlığı örnekler ile açıklanmıştır.

Anahtar Kelimeler: Bilgisayarlı Aktarma, Türk Dilleri, Türkçe Gramer,

I. **Introduction**

[♣] Istanbul Bilgi University, Computer Engineering Department, Faculty of Engineering and Natural Sciences, murat.orhun@bilgi.edu.tr , ORCID NO:

Makalenin Dergiye Ulaşma Tarihi:08.08.2018 Yayın Kabul Tarihi: 07.09.2018

Turkic languages include a lot of languages such as Turkish, Azerbaijani, Uzbek, Turkmen, Kazakh, Kyrgyz and Uyghur etc. Turkic languages are belonging to the Ural- Altaic language family. All the Turkic languages are agglutinative languages which have productive inflectional and derivational morphology. It means a new word could be generated just adding a simple suffix to end of a word or putting a prefix in front of a rood word. In Turkic languages, it is very common that one word could be followed more than one suffix. In this way, new words could be generated form the one root word. For machine translation, it is necessary to keep a bilingual dictionary to translate related words.

But it is not possible or useful to keep all words in a dictionary that generated from root words with adding suffixes. For example, the Turkish word "iş" (work) the root or base of following words that have different meaning as below:

iş	work
iş+im	my work
iş+i	her/his work
iş+imiz	our work
iş+in	your work
iş+ler	works, tasks
iş+çi	worker
iş+çi+si	worker of some one, worker of some kind
iş+çi+ler	workers
iş+çi+ler+e	to workers
iş+çi+ler+den	from workers
iş+çi+ler+in	of the workers, workers'
iş+çi+ler+in+ki+dek	same as of workers
iş+çi+ler+in+ki+dek+çe+sene	looks like that belong to workers
iş+siz	jobless
iş+le	do work
iş+le+me	do not work
iş+lem	computational
iş+lek	busy
iş+lem+ci	processor
iş+lev	function
iş+lev+sel	functional

In this example, the root of all words is “iş”, after the boundary sign, marked with the “+” character, different suffix can be followed. Some words followed by only single suffix, but the number of suffix can be increased according to grammar of the language. The structure of the words almost similar in that example and they have the same root words but suffixes are different. On the other hand, the structure of the English translation are quite different.

The longest word in the example is “iş+çi+ler+in+ki+dek+çe+sene” and this word can be followed more suffix according to Turkish language grammar and much more longer words could be created (Oflazer 1994, Banguoğlu 2000). Whenever a suffix is attached, the mean and category of the word also changes in parallel. For example, the word “iş” (work) is belong to the noun category of the Turkish language. When the “çi” is followed, a new noun “worker” is created and it is a person. If the suffix “le” is followed by “iş”, an imperative verb (do work) is created. In case, the suffix “siz” is followed the root word, adjective (jobless) is created. Whenever property of a word changes, it affects structure of a sentence. Therefore, it is crucial to analyze all information of a word in natural language processing, especially fields such as in machine translation applications. The same kind of examples could be found almost in all Turkic languages. Though different alphabets are used officially in different Turkics languages, but Latin scripts are used in this article to compare related Turkic alphabets. For Uyghur languages, the common Uyghur Latin-Script (ULY) is preferred (Janbaz et al. 2006).

For example: it is possible to write an Uyghur word as follows according to Uyghur language grammar (Kaşgralı 1992, Osmanof 1997, Tömür 2003, Tehür 2010).

UYGHURLARNINGKIDEKMISH

The mean of this word is, “things that belong to the Uyghur people” and this word can be splittable into the following morphological information as below:

UYGHUR+LAR+NING+KI+DEK+MISH

Root of the word is, “Uyghur” and rest of them are different suffix. This word also can be extended with more suffixes. For example:

UYGHUR+LAR+NING+KI+DEK+MISH+SE

The following Uyghur word is created with adding different suffixes to the root word “ITTIPAQ” and it means “They might have been trying to divide people”. This word also could be split into morphemes according to Uyghur grammar.

ITTIPAQSIZLASHTURALMAYWATQANLIQINGLARDINMIKINTANG

Of course in modern Uyghur language there is not such a long word, but this words is grammatically correct according to the language grammar and could be analyzed with a well defined morphological analyzer. In general, word order and syntactic structure of all Turkic language are similar (Karahan et al. 2013).

The general order of the sentence is “ Subject + Object + Verb”. For example:

Turkish:	Ben okula gideceğim.
Uyghur:	Men mektepke barimen.
Uzbek:	Men maktabga boraman.
Kazakh:	Mén mektepke baramin.
Kyrgyz:	Mén méktepke baram.
Azerbaijani:	Men mektebe gederem.

These sentences could be translated into English such as “I am going to school”. Mean while in all Turkic language have the same word order, even there are same number of words exist. This is a simple example just show to word order similarity of the some Turkic languages. Even all Turkic language have such similarity, but there are same critic differences between them as explained in (Orhun et al. 2009a). In natural language processing especially in machine translations, it the first step to analyze the words. Therefore morphological analyzing is the core point of the natural language processing. For machine translation, if there are same tags used to keep the morphological information of the words, then it is easy to implement a translations system between different languages.

In this paper there are some of the morphological analyzers of Turkic languages have been analyzed and explained with examples. In the meantime we suggest the possibility that we can extent the existing systems to analyze other languages. This paper is organized as follows: there are some machine translation and morphological analyzers of Turkic languages are introduced in section 2. In section 3, the two-level morphology and two-level rules have been explained with Uyghur language as an example. In section 4, a simple word translation between Uyghur and Turkish is given. The conclusion is given in the last section.

II. Machine Translation and Morphological Analyzers of Turkic Languages

The first machine translation system that implemented for Turkic languages is the system that translates from Turkish to Azerbaijani language (Hamzaoğlu 1993). In this system words are translated according to root words. Therefore there is a Turkish - Azerbaijani bilingual dictionary is used. In order to select synonym words, some extra tables used with word explanation. Following this Turkish – Azerbaijani translation system, another translation system implemented that translates form from Turkish to Crimean Tatar (Altıntaş 2000), which works on simple sentences without solving ambiguities. The latest translation system implemented for Turkic language is a bilingual translation between Turkmen and Turkish languages (Tantuğ 2007). This system is the best translation system that implemented for Turkic languages so far (Tantuğ et al. 2007). Hybrid methods used such at different level of language models to select most suitable words among words that have similar or close meanings (Figure 1). For example, the Turkmen word “adam” (person, man) could be translated in to Turkish such as “adam” or “insan”. Both of these translation could be accepted. But to select the most suitable word, in mean of sentence structure or fluidity, related language modules should be considered. Turkmen word “geple” (speak) could be translated in to Turkish such as “konuş” or söyle”.

Even these two Turkish words have the same meaning, but sometimes they can not be used in place of each other.

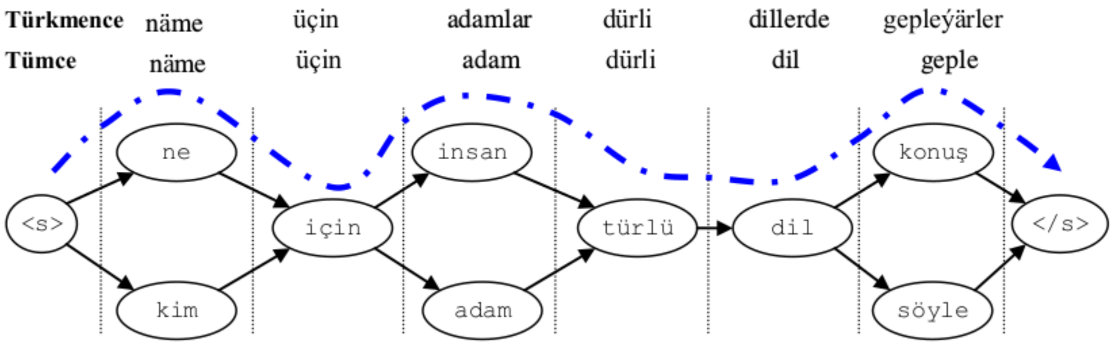

Figure 1: Machine translation system from Turkmen to Turksih (Tantuğ et al. 2007).

There is a machine translation between Turkic and non-Turkic languages, Uyghur to Japanese translations, has high performance (Mahsut et al. 2004). Because of Uyghur and Japanese have the same grammatical property and the same sentence structure, word orders of the translated sentence kept with same order of the source sentence. In order to solve ambiguities, some rules defined to translate words that has multiple meaning. This system used a dictionary that has 22,000 words and performance of the system has been evaluated about 85%. There are some new machine translation systems such as, Turkish to Uyghur and Uyghur to Uzbek are still in progress.

In general, all machine translation systems that implemented for Turkic languages have common properties. All of them translates from other Turkic languages to Turkey Turkish and all Turkic language have morphological analyzers. Though all of those languages closely related each other, it is impossible use a single morphologic analyzer to all different language directly. Therefore, a specific morphologic analyzer developed for Turkish (Oflazer 1994), Crimean Tatar (Altintaş et al. 2001), Turkmen (Tantuğ et al. 2006) and Uyghur (Orhun et al. 2008, Orhun et al. 2009b) languages. Recently morphological analyzers have been implemented for following Turkic language such as Qazan Tatar (Gökgöz et al. 2011), Kazakh (Zafer et al. 2011, Kessikbayeva et al. 2014), Kyrgyz (Görmez et al. 2011, Washington et al. 2012) and Uzbek (Matlatipov et al. 2009) etc. Except the Uzbek language morphological analyzer, rest of the morphological analyzers have been implemented with two-level morphology concept (Sproat 1992) and using finite-state transducer tools (Karttunen 1997, Karttunen 1983)

In machine translation, it is important to translate source sentence contents into target language without changing original mean. Therefore the quality of translation affected by both of dictionary and sentence structure. There are so many common words among all Turkic languages. For example there are 7000 common words have been listed for Turkish, Azerbaijani, Bashqurt, Kazakh, Kyrgyz, Uzbek, Tatar, Turkmen and Uyghur languages (Ercilasun et al. 1991). Languages that belong to the same groups of Turkic family, there are even more common words. For example, Turkish-Azerbaijani-Türkmen, Uzbek-Uyghur, Kazakh-Kyrgyz languages pairs have so many similar words. Some linguists assume that Uyghur and Uzbek are not two different language, on the other hand , they accept them just a two different dialect of the same language.

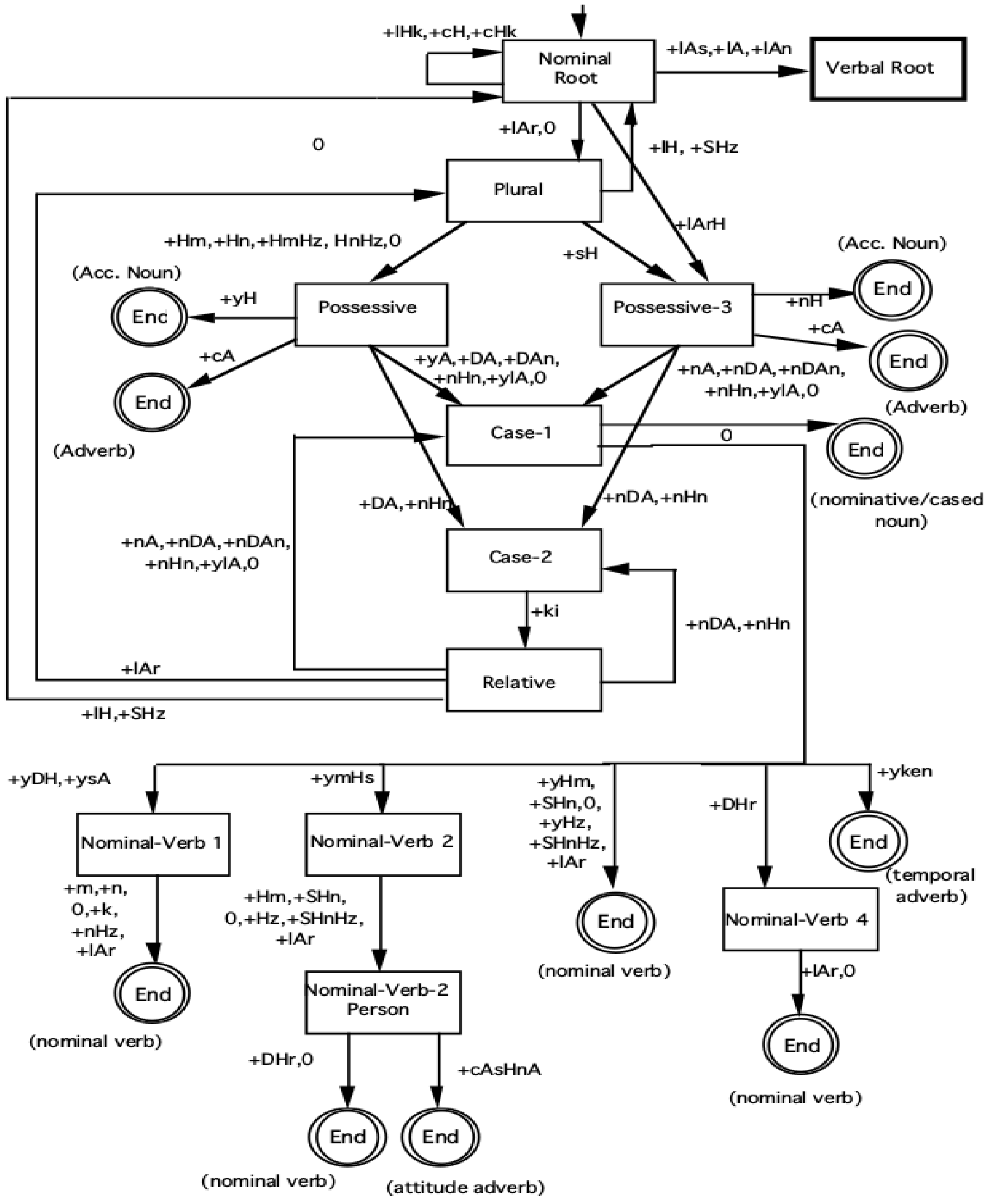

Figure 2: Two-level morphological analyzer of Turkish nouns (Oflazer 1994).

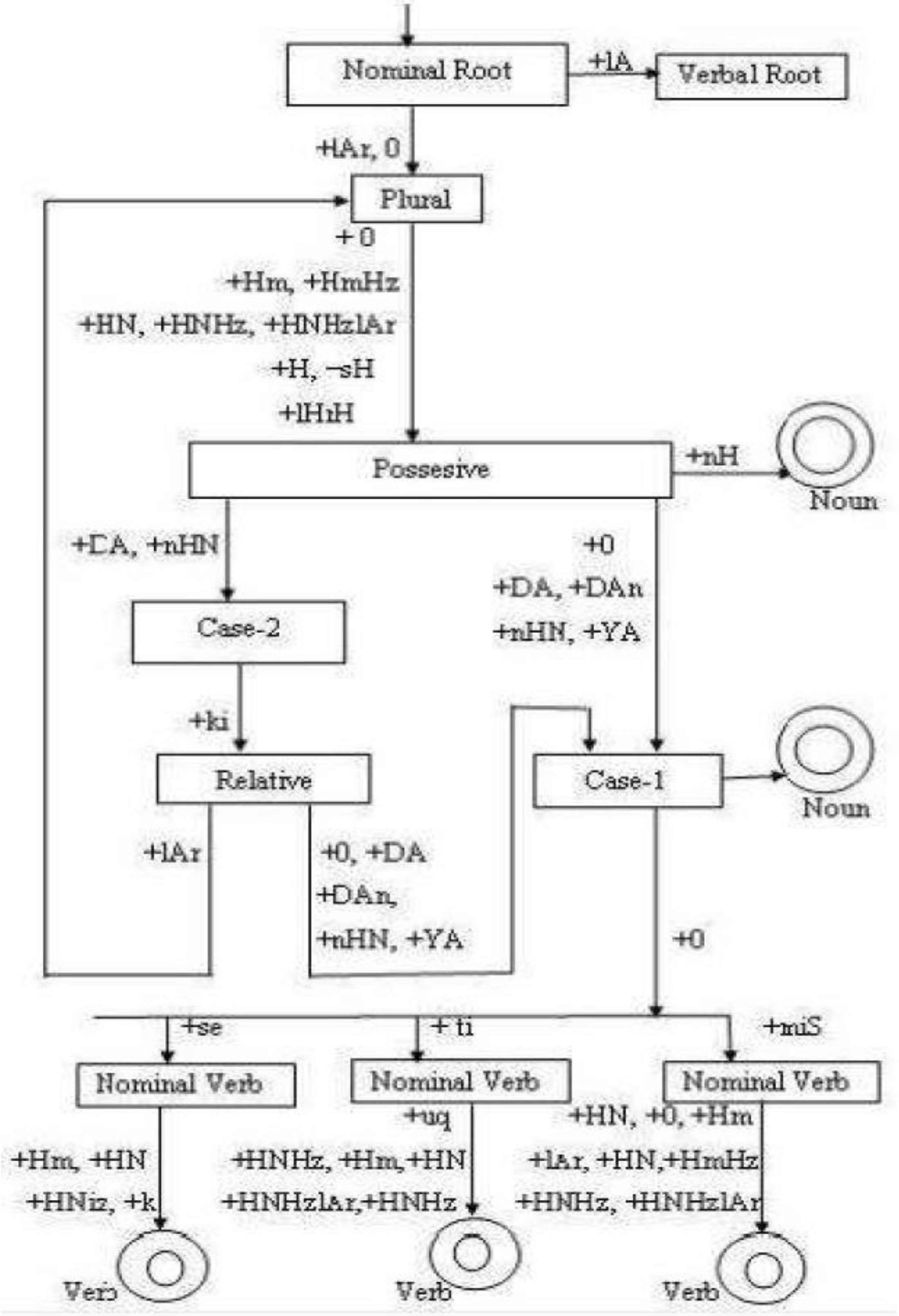

Figure 3: Two-level morphological analyzer of Uyghur nouns (Orhun 2008)

In translation, to form the sentence structure of the target sentence is most important and most difficult task. Because all Turkic languages have almost the same sentence structure and have the same word order in sentence, it takes less effort implement translation systems for Turkic languages.

Turkish is the most computational researched language among all Turkic languages.

There are so many important projects have been done so far. For example, full functional morphological analyzer, tagged corpus, content summarization, entity recognition, solving disambiguation among words etc. Though Turkish and other Turkic languages similar and have common properties, research result of a language cannot be applied on other language directly. Because there are still some difference between different languages that can not be neglected. Therefore other Turkic languages may have to be researched based on Turkish.

Actually all two-level morphological analyzers that implemented so far are have been implemented based on Turkish morphological analyzer as showed in Figure 2. For example the morphological analyzer for Uyghur nouns (Figure 3) almost same to the Turkish noun analyzer. In Figure 2 and Figure 3, uppercase characters represents a character that harmonizes according to some conditions. Lower case characters represents they are none changing characters. For example, in both figure there is a suffix with “lAr” characters. This is a plural suffix in all Turkic languages and the uppercase /A/ character will be replaced with /a/ or /e/ according to followed vowels.

In Figure 2 and Figure 3, the rectangles define middle states of a word and from these states a new suffix could be added and go to a new state. The double circles define final states. For example, in Uyghur language, if a word followed by the accusative suffix “ni”, then this word reaches the final state. Because there are not suffixed allowed to follow the accusative suffix according to language grammar. Rest of the characters of the two figure have been explained in related papers in detail (Oflazer 1994, Orhun 2008).

As a result, there are two kinds of characters analyzed in this morphological analyzer. On kind of characters are real characters that carry proper information and they called surface level characters. Second kind of characters, such as uppercase, they exit only at lexical level any they change according to language grammar. Lexical level characters always controls surface level characters and all language related rules have been applied at the lexical level. Because of this reason, this kind of morphological analyzers called two-level (lexical – surface) morphological analyzer.

Two-level morphological analyzer is one of the most commonly used technique for analyzing language. Especially for agglutinative languages, very complected problems could be solved withs some simple rules. Most of the morphological analyzers, even not agglutinative languages, have implemented with two-level transducers. For example, Japanese (Alam 1983), English (Karttunen 1983), Finnish (Koskenniemi 1985), Romanian (Khan 2016) etc.

III The Uyghur Language and Two-Level Rules

Alphabet is one of the most important topics for machine translations. In Turkic languages, there is 20-30 percentage words are the same in average (Kaşgarlı 1992). Even there are 60-70 percentage word are the same in Uzbek and Uyghur languages. For machine translations this is a big advantage. Unfortunately, different alphabets have been used for different Turkic languages. For example, some of the former Soviet Republics such as Kirghizstan and Kazakhstan (Kazakhstan is planning to use Latin Script) are using Cyrillic alphabet, while some of the other republics such as Turkmenistan, Azerbaijan and Uzbekistan are using Latin alphabet. Even Uzbekistan has announced that Latin alphabet be official script of the county, still most of the official documents are printed with Cyrillic alphabet. In this paper we use the Uyghur Latin alphabet as described in (Janbaz et al. 2006) and explain how could be the alphabet applied to the morphologic analyzers that could be used with XEROX finite state tools (Karttunen 1997).

Uyghurs live mainly Sinking Uyghur autonomy region in China and they use Arabic alphabets officially. Also they live in central Asian republics and use Cyrillic alphabet. Very few Uyghurs live in western world and they use Latin-Script. In order to communicate between these three groups, the Uyghur Computer Science Association (UKIJ) has implemented a alphabet translation software and distributed with free of charge.

The contemporary Uyghur alphabet is composed of 32 characters and they are categorized into two groups such as vowels and constants. There are 8 vowels (a e é i o ö u ü) and 24 consonants (b p t j ch x d r z j s sh gh f q k g ng l m n h w y). In Uyghur language all words consists of with at least with one vowel and vowels are base of all words [36]. Because some not ASCII characters used to represent some of Uyghur character in Latin script such as, /ö/, /ü/, /é/, these characters have been represent with uppercase characters at lexical level in order to apply two-level rules correctly (Orhun et al. 2008, Karttunen 1997).

The XEROX tools works with the standard ASCII characters and it is important to modify some not standard both of Uyghur and Turkish characters that consists of double characters (Orhun et al. 2008). To solve such as not ASCII character problem, standard characters have been represented with lowercase characters and non ASCII characters have been represented with uppercase characters. For example double charactered not standard characters such as /ch/, /sh/, /gh/, /zh/ and /ng/ are represented with /C/, /S/, /G/, /Z/ and /N/.

In generated, two-level tools work completely independent of any natural language. After alphabets and rules of a language are defined and submitted, these tools works according to defined rules and analyze words of a language. As a result, two-level tools doesn’t understand any language, what they are understands are alphabets and related language rules.

Because of different languages have different language properties, they need to be analyzed according to related language grammar. To defined general rules for common properties of a language, some characters are classified as sets. For example, in Uyghur Language, vowels are classified three sub categorize such as front vowel, back vowel and middle vowels (Orhun et al. 2008).

Front vowel: a , e, o

Back vowel: ö, u, ü

Middle vowel: é, i

To write more specific rules, it may be possible to define more sets or categories for characters. For example, according to location of tongue, length of sound etc. In the same way for constants, there are 5 different rules have been defined for Uyghur (Orhun et al. 2008). In all Turkic languages, a plural words could be created with adding the “lar” or “ler” suffix after a noun. To add the “lar” or “ler”, will be decided according to the vowel in the last syllable of a word. In case the vowel of the last syllable belong to the front vowel, then “lar” should be added, other wise “ler” will be added.

In Uyghur there are 8 vowels that three of them belong to front vowel category while another three of them belong to back vowel. Also there are two vowel that belong to middle vowel category. If the vowel of the last syllable belong to the middle vowel category, it is not so clear to add “lar” or “ler”. In this case, it is necessary to check other vowels. Even for some words, it is impossible to apply a general rule. This situation is appears for adapted words from other languages.

To write efficient rules, for example such as “lar” and “ler” suffix. Some changing characters represented with a variable characters at the lexical level. If the “lar” and “ler” suffixes are compared, they different from only one character, /a/ and /e/. There fore if the /a/ and /e/ characters represented with are meta character, for example such as /A/, then the plural suffix for Uyghur , even for all Turkic languages is, represented as “lAr”. It means there are only one suffix for plural for all Turkic languages. The /A/ character in the suffix “lAr”, will change according to vowel in a word that it follows. In the same way, it is possible to define meta characters for dative and locative suffix for other Turkic languages. For example, it is possible to define some meta characters for Uyghur language as below (Orhun et al. 2008):

Lexical Meta Vowels:

A = a, e

H = i, u, U

Lexical Meta Consonants:

D = d, t

Y = G, q, g, k

After such character sets and meta characters have been defined, it is possible to write two level rule as below. As mentions in previous section, these rules works at the lexical level and all surface (true) characters are controlled by these rules. The following rules explains how to convert the lexical meta character /A/ into /a/ or /e/.

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

[CONSBV+][:MIDV][:CONSBV]*]+(%+:0)[:CONSBV|CONSBV:]*_

2.A:e=>[[[CONSFV][:MIDV][CONS]*]+][[MIDV][CONS]*]*(%+:0)[:CONS|CONS:]* _

[[:CONS]*[:FRONTV][CONS]*]+(%+:0)[:CONS |CONS:]* _

[[CONS]*[:MIDV][CONSFV]+]+(%+:0)[:CONS |CONS:]* _

After these two level rules have been defined, we can add plural suffixes add lexical level as bellow. For example:

Lexical: kitab+lAr

Surface: kitab0lar

Real: kitaplar // the /0/ is not displayed in real word

The root of the word is “kitab” (book) and the vowel in the syllable is /a/. The /a/ vowel is belong top the front vowel category. Therefor the /A/ should be converted into /a/. When two-level tool is working, it checks the vowel in the last syllable and finds the /a/ character. The it checks which set included the /a/ character. Then it decides to convert /A/ into /a/ or /e/. While two-level tools are converting lexical characters into surface characters, the convert the /+/ sign into the special character /0/. The /0/ character never displayed on the surface of a word.

The following example shows how the meta /A/ changes in /e/ character after the “qeGez” (papar).

Lexical: qeGez+lAr

Surface: qeGez0ler

Because of the vowel in the last syllable of the word is belong to the back vowel, the meta /A/ character should be changed into the /e/ character.

IV. WORD TRANSLATION BETWEEN UYGHUR AND TURKISH LANGUAGE

After the morphologic analyzer defined, it is simple to define legal words, as called morphotactic (Sproat 1992). We define a simple lexicon for Uygur nouns which followed plural and case suffix which depending on the Figure 3.

The extensive lexicons are given in (Orhun et al. 2009b, Orhun et al. 2008).

Lexicon structure for nouns	Explanation of tags

```
LEXICON Noun
   kitap NounPOS;
   kelem NounPOS;

LEXICON NounPOS
  +Noun:0 Suffixes;

LEXICON Suffixes
 +A3pl:lAr Noun-Plural;
 +A3sg:0 Noun-Plural;
LEXICON Noun-Plural
 +Pnon:0 Noun-Posessive;
 +P1sg:Hm Noun-Posessive;
 +P1pl:HmHz Noun-Posessive;
 ……………….

LEXICON Noun-Possessive
   +Nom: 0 Final;
   +Acc: nH Final;
   +Loc: DA Case;
   +Dat: YA Case;
   ……………………….

LEXICON Case
   ^DB+Adj+Rel: ki Final;
    ………………………….
LEXICON Final
   #;
```

+Noun -> Nouns

+A3sg -> 3-person Singular (Number-Person agreement)

+A3pl -> 3-person Plural (Number-Person agreement)

+Pnon -> Pronoun (no agreement)

+P1sg ->1-perons singular (Possessive agreement)

+P1sg ->2-perons singular (Possessive agreement)

+Acc -> Accusative

+Dat -> Dative

+Loc -> Locative

+Adj -> Adjective

+Nom -> Nominative

^DB -> Derivational Boundary

After we have defined these tags, we can get morphologic information with morphological analyzer as bellow. For example:

Surface Level: kitab

Lexical Level: kitab+0

Morphologic Results: kitab+Noun+ A3sg+ Pnon+ Nom

It means, the root of the word is, “kitab”, it is a noun,3rd person single, no agreement with persons, and nominative.

Surface Level: kitablar

Lexical Level: kitab+lAr

Morphologic Results: kitap+Noun+ A3pl+ Pnon+ Nom

It means, the root of the word is, “kitab”, it is a noun, 3rd person plural, no agreement with persons, and nominative.

Surface Level: kitablarni

Lexical Level: kitab+lAr+nH

Morphologic Results: kitab+Noun+ A3pl+ Pnon+ Acc

It means, the root of the word is, “kitab”, it is a noun, 3rd person plural, no agreement with persons ,and in accusative case.

Surface Level: kitablarimizni

Lexical Level: kitab+lAr+HmHz+nH

Morphologic Results: kitab+Noun+ A3pl+ P1pl+ Acc

It means, the root of the word is, “kitab”, it is a noun, 1-person plural ,1-person plural possessive, and in Accusative case.

If the lexicon includes all of the words in a language, then its morphologic properties could be analyzed in the same way. If the analyzed words is not found in the lexicon, it means it is not valid word in that languages. With this property of lexicon, it is possible to spell check of a document.

Now let’s analyze a word in Turkish with the morphological analyzer for Turkish (Oflazer 1994) and translate the word, “evlerimizde” means in English is “in our houses”, to Uyghur language.

TR: Surface Level: evlerimizde

TR: Lexical Level: ev+lAr+HmHz+ DA

TR: Morphological Results: ev+Noun+A3pl+P1pl+Loc

It means, the root of the word is “ev”, it is a noun,1-person, plural,1-person plural possessive, in the accusative case. We can translate the root word “ev” in Turkish to Uyghur language as “Oy” with a bilingual dictionary and apply the morphological results of the Turkish word “ev” to the Uyghur word “Oy”.

TR: Surface Level: evlerimizde

TR: Lexical Level: ev+lAr+HmHz+ DA

TR: Morphological Results: ev+Noun+A3pl+P1pl+Loc

UYG: Morphological Results: Oy+Noun+A3pl+P1pl+Loc

UYG: Lexical Level: Oy+Noun+lAr+HmHz+ DA

UYG: Surface Level: Oylirimizde

In the same way, if we want to translate the Uyghur word “Oylirimizde” to Turkish, the system work in the same way and get the correct translation. Because, the Two-Level morphology works in the two directions and XEROX (Karttunen 1997) implemented very useful tools for two level morphologic analyzing. If we submit a word to surface, we can get it is lexical level and morphological information. Also, if we have some morphological information of a word, then we can get its lexical form and get its surface.

Meanwhile, there are other tools available such as KIMMO and ZEMBEREK (Karttunen 1983). Here ZEMBEREK is an open sourced tool for language processing that implemented in Java. This tool could be downloaded and improved for further researches. It is also one of the main tools of the Pardus.

V. CONCLUSION

In this article, most recent researches explained about machine translation systems for Turkic languages. Also these systems summarized with Turkish and Uyghur translation system such an example. The core technique of all systems are that they are based on two-level morphology. Actually, two-level morphology is a common method for analyzing morphology of a language and it is independent of any language. A lot of most computational researched language such as Japanese, English, Finnish and Romanian are analyzed with this technique (Karttunen 1997).

From the example that explained in the last section, if the same tags are used for keeping morphologic information, it is possible to implement a machine translations system more effectively for different languages. Because of all Turkic language have common property it is possible to define common lexical rules for all language. For special cases, may be special rules could be defined. In general verbs in all Turkic languages are very complicated even they have common properties, still there are difference exist that can not be neglected. For example there are different cases exist of same words that could be analyzed with computers only (Eziz 2007, Belikiz 2007).

By nature, it is difficult to define rules for all cases for a natural language. Therefore the most effective solution is make common corpus all Turkic languages. In this case, with using the common corpus, it is possible to evaluate words and check translation quality more effectively (Hakkani-Tür 2000, Belikiz 2004). In the last decades, Turkic language speaking countries reforming their alphabets and sated that they are going to use Latin alphabets. This is a good progress for making a natural corpus and write a common machine translation system for all Turkic languages.

REFERENCES

OFLAZER, Kemal (1994). "Two-Level Description of Turkish Morphology". ***Literary and Linguistic*** *Computing*, Vol. 9, No:2

ALAM, Y. Sasaki (1983). "A Two-Level Morphological Analysis of Japanese". **Texas Linguistics Forum,** 22:229-252.

ALTINTAS, Kemal et al. (2001). "A Morphological Analyzer for Crimean Tatar". **Proceedings of the 10th Turkish Symposium on Artificial Intelligence and Neural** Networks .

ALTINTAŞ, Kemal (2000). *Turkish to Crimean Tatar Machine Translation System*. Msc Thesis. Ankara: Bilkent University.

BANGUOĞLU, Tahsin (2000). ***Türkçenin Grameri*: ** Türk Dil Kurumu.

BELIKIZ (2004). "Corpus Bases Word-Class Taggin of Uyghur". **The 5th Xinjiang Youth Academic Conference.** Urumqi, China.

BELIKIZ (2007). "The 3253 different word forms Uygur Verb "qil" ". Corpus Linguistics and Corpus Based Reseach. Department of Linguistics, College of Anthropology. Xinjiang Normal University, Xinjiang, China.

ERCILASUN B. Ahmet et al. (1991). ***Karşılaştırmalı Türk Lehçeleri Sözlüğü I*,** Kültür Bakanlığı Yayınları, Ankara.

EZIZ, Gülnar (2007). "Resistance to Borrowing of Uyghur Verbs". **Annual Conference. University of Washington**, October18-21, (2007).

GÖKGÖZ, Ercan et al. (2011). "Two-level Qazan Tatar Morphology". **The 1st International Conference on Foreign Language Teaching and Applied Linguistics,** Sarajevo, Bosnia. May 5-7.

GÖRMEZ, Zeliha et al. (2011). "An overview of Two-level Finite State KyrgyzMorphology". **The 2nd International Symposium on Computing in Science & Engineering.** Kuşadası, Aydın, Turkey. 1-4 June.

Hakkini-Tür, Dilek (2000). *Statistical Modeling of Agglutinative Languages.* PhD Thesis. Ankara: Bilkent University.

HAMZAOĞLU, İlyas (1993). *Machine translation from Turkish to other Turkic languages and an implementation for the Azeri languages*, **Institute for Graduate Studies in Science and Engineering.** MSc Thesis. İstanbul: Bogazici University.

JANBAZ, A. Waris et al. (2006). "An Introduction to Latin-Script Uyghur". Middle East &

Central Asia Politics, **Economics, and Society Conference**. Sept 7-9, University of Utah, Salt Lake City, USA .

KARAHAN, Leyla et al. (2013). ***Karşılaştırmalı Türk Lehçeleri Grameri I.*** TDK yayınları. Ankara.

KARTTUNEN, Lauri (1983). KIMMO: *A General Morphological Processor,* **in Texas Linguistic Forum**, Texas, USA, (1983).

KARTTUNEN, Lauri et al. (1983). K.Wittenburg, "A Two-Level Morphological Analysis of English". **Texas Linguistics Forum**, 22:217-228.

KARTTUNEN, Lauri et al. (1997). *Xerox Finite State Tool*. **Technical Report,** Xerox Research Centre, Europe.

KAŞGARLI, S. Mehmet (1992). ***Modern Uygur Türkçesi Grameri*,** İstanbul.

KESSIKABAYEVA, Gulshat et al. (2014). "Rule Based Morphological Analyzer of Kazakh". **Language Proceedings** of the 2014 Joint Meeting of SIGMORPHON and SIGFSM. Baltimore, Maryland, USA, pp: 46–54.

KHAN, R (2016). "A Two-Level Morphological Analysis of Romanian", In **TexasLinguistic Forum,** Texas, USA, pp.253-270, (1983).

KOSKENNIEMI, Kimmo (1985). "An Application of the Two-Level Model to Finnish". In Fred Karlsson,editor, **Computational Morphosyntax, a report on research 1981-1984. University of Helsinki Department of General Linguistics**.

MAHSUT, Muhtar et al. (2004). "An Experimention Japanese-Uighur Machine Translation and Its Evaluation. **AMTA 2004**, LNAI 3265, pp.208-216.

MATLATIPOV, Gayrat et al. (2009). "Representation of Uzbek Morphology in Prolog", **Aspects of Natural Language Processing, Springer-Verlag**, Berlin. p:83-110

ORHUN, Murat et al. (2008). "Rule Based Analysis of the Uyghur Nouns". **Proceedings of the International Conference on Asian Language Processing** (IALP). Chiang Mai, Thailand, 12-14 November.

ORHUN, Murat et al. (2009a). "Computational comparison of the Uyghur and Turkish Grammar". **The 2nd IEEE International Conference on Computer Science and Information Technology.** Beijing, China. 8-11 August.

ORHUN, Murat et al. (2009b). "Rule Based Tagging of the Uyghur Verbs". Fourth International Conference on Intelligent Computing and Information Systems. Faculty of Computer & Information Science. Ain Shams University Cairo, Egypt, 19-22, March.

OSMANOF, Mirsultan (1997). ***Hazirqi Zaman Uyghur Edebiy Tilining İmla ve Teleppuz*** *Lughiti.* Xin Jiang Xeliq Neshiryatı. Yanwar.

SPROAT, Richard (1992). ***Morphology and Computation****.* MIT Press.

TANTUĞ, A. Cüneyd et al. (2006). "Computer Analysis of the Turkmen Language Morphology". Fin-TAL, **Lecture Notes in Computer Science**. 4139:186-193.

TANTUĞ, A.Cüneyd (2007). *A Hybird Model for Machine Translation Betwee*

Agglutinative and Related Languages. PhD Thesis. Istanbul: Istanbul Technical University.

TANTUĞ, A.Cüneyd et al. (2007). “Machine Translation between Turkic Languages”.Proceedings of the ACL 2007 Demo and Poster Sessions. p189–192, Prague.

TEHÜR, Y. Abdulla et al. (2010). ***Hazirqi Zaman Uyghur Dili****.* Xin Jiang Halq Neshiryati, Urumqi, China.

TÖMÜR, Hamit (2003). *Modern Uygur Grammar (Morphology).* Yildiz Teknik Üniversitesi, Fen-Ed. Fak. T.D.E Bölümü. Istanbul, Türkiye.

WASHINGTON, D. Johathan et al. (2012). “A Finite-state morphological transducer for Kyrgyz”. **Proceedings of the Eight International Conference on Language Resources and Evaluation** (LREC'12). Istanbul, Turkey, 23-25, May.

ZAFER, H. Regit et al. (2011). “Two-level Description of Kazakh Morphology”. **The 1st International Conference on Foreign Language Teaching and Applied Linguistics.** Sarajevo, Bosnia. May 5-7.