1. INTRODUCTION

This paper describes identifying or defining Uyghur words according to their syllabic properties. Uyghur is a Turkic language spoken mainly in Xinjiang Uyghur autonomous region in China. By morphological structures, all Uyghur words have standard syllabic properties and all words can be split into syllables by applying general syllabic rules [1]. But Uyghur language is one of the oldest language in the Turkic language family and it is spoken in wide geographic region and counties such as Uyghur autonomous region in China, Afghanistan, Kazakhstan, Kyrgyzstan, Uzbekistan, Turkey, USA and some European countries. It includes many words that are not of Uyghur origin [1]. Most Uyghur speakers live in the Uyghur autonomous region in China and the contemporary Uyghur language is heavily affected by Chinese words. There are also a lot of words adopted from Russian in Central Asian republics. In addition, the Uyghur language is also affected by Arabic and Persian words because of religion and geographic relations. Therefore, to study or analyze Uyghur language with computer based methods, it is necessary to define the origin of a word properly. In natural language studies, the alphabet is one of the most important factors. A range of alphabets and different numbers of characters have been used in different part of the world to write the Uyghur language. For example, the Arabic based alphabet is used in China (Figure 1), while the Cyrillic alphabet is used in Central Asian republics (Figure 2) and the Latin based alphabet is used in western countries (Figure 3).

Figure 1: Arabic Alphabet [2]
Therefore, to study the Uyghur language, it is necessary to study the relationship between these different alphabets. Characters used in an alphabet directly affect word structures and spelling rules. For example, in Central Asian republics, there are some Russian characters have been used to write Russian adapted words, but a single Russian character can represent two characters in the Uyghur Language. To study the Uyghur language as a single unit, it is important to implement correct the algorithm to convert one alphabet into another. Even though the Arabic based Uyghur alphabet is the official alphabet in Sin Kiang Uyghur autonomous region, the Latin based alphabet is commonly used. In this paper Uyghur words are split according to the Latin based alphabet adopted by the UKIJ [3-4] (Figure 4).

This paper describes mainly how to identify or define Uyghur native words according to their syllabic properties. Comparing the difference between different Turkic languages or none Turkic languages syllabic properties is out of the scope of this paper.

This paper is organized as follows: after providing short information about NLP and Uyghur language in the first section, syllabic and morphological properties of Uyghur words have been explained in the second section. The third section describes implementation of the algorithm that splits words into syllables and in the last section the algorithm has been evaluated and the result has been explained.

2. SYLLABIC and MORPHOLOGICAL PROPERTIES OF UYGHUR WORDS

To study syllabic properties of words, the first thing to do is analyze morphologic properties of those words. Uyghur is an agglutinative language with word structures formed by productive affixations of derivational and inflectional suffixes to root words. For example:

**SHEHIRDEKILERNINGKIMISHDEK**

Which can be broken down into morphemes as follows:

**SHEHIR+DE+KI+LER+NING+MISH+DEK**

Where the “+” indicates morpheme boundaries. This word can be translated into English such as “as if they belong to a city”. The root of this words is “SHEHIR” and rest of the morphemes add external meaning to the root word. Whenever a new morpheme is affixed, a new category is created. While a new morpheme or suffix is affixed, vowels in a morpheme have to agree with the preceding vowel in certain aspects to achieve vowel harmony, although there are small number of exceptions. In some cases, vowels changed or deleted from the root words [1, 12]. Similarly, such modifications appear about consonants in root word and affixed morphemes.

Complicated morphological structures of a word, especially agglutinative languages, make it more complicated to study morphological, lexical and syntactic property of a language.

Uyghur origin or native words and adapted words have different morphological structures, therefore some computational morphological analyzers cannot solve all Uyghur words correctly [13-14]. If a word is identified, before it is analyzed and a none Uyghur origin word is elected. Next, different methods of analysis are suggested and the performance of the morphological analysis may be improved. In natural language processing, for agglutinative languages, a morphological analyzer is the most important part and it provides the fundamentals for further research.

---

**Figure 2:** Cyrillic Alphabet [2]

**Figure 3:** Latin Alphabet [3]

**Figure 4:** Latin-Arabic Conversion Table [4]
In contemporary Uyghur language there are eight vowels and 24 consonants.

In general, words used in the contemporary Uyghur language can be analyzed according to rules that describe above six standard rules [1].

According to those rules, an Uyghur word may consist of a single syllable or unlimited (with affixed suffixes) numbers of syllables. To correctly find out syllables in a word, it is important to define borders of syllables. In general, the syllable borders can be defined according to the following rules [1]. There are some special cases that do not follow these rules though these cases not included in this paper.

1. If there is a consonant between two vowels, this consonant must be in the same syllable with the first vowel. For example:
   yürek->yü+rek, ana - >a+na, yemekxana->ye+mek+ha+na

2. If there are two consonants between two vowels, first consonant must be put in a syllable with the first vowel, and next consonant belongs to second syllable with the last vowel. The main rule applied here is, in a syllable, only one vowel is existin a Uyghur origin word. For example:
   mektep ->mek+t+ep, ketmen->ket+men
   baghwen ->bagh+w+en

3. If there are three consonants between two vowels, the first two consonants must put in the first syllable with the firs vowel, and the last consonant and the last vowel grouped a syllable. For example:
   gherptin->gherp+t+in, shergshunas->sherg+shunas
   gherpliq->gherp+liq.

4. If the last character of a word is consonant, and a suffix with first character vowel is added, then the last consonant of the word makes a syllable with the added vowel character. For example:
   kitab+i1 ->ki+t+ka, qadir+i1 ->qa+di+ri

In Uyghur language, there are some special cases to define the border of a syllable, and it is dependent on vowel harmonization. When a vowel is changed, it also affects some consonants and these changes affect the border of the syllable. But such very specific cases are not included in this paper.

In some cases, a well implemented morphological analyzer can be used as a syllable splitter, but a morphological
3. IMPLEMENTATION of the ALGORITHM

To implement the algorithm that splits Uyghur words into syllables, both Cyrillic and Arabic characters have been converted into Latin characters according to the source of the files. After that, the six rules have been applied on all words. In the last step, adjust syllable borders and decide if the created syllables map the standard Uyghur syllable styles or not. The algorithm that splits words into syllables can be represented as in Figure 5.

Figure 5: Word syllabling algorithm

If any words cannot be split, those words are considered as adopted words from other languages. To analyze these adopted words, alternative methods could be suggested. It may be open topic for this kind of problem. Because there are many adopted words from different languages.

As shown in Figure 5, there are two main parts of this algorithm, splitting into syllables and adjusting syllable borders. When the borders are adjusted, the number of characters in a syllable may be changed. In general, the maximum number of characters in a syllable is four and the minimum number of characters is one (one vowel).

4. RESULTS AND DISCUSSION

This algorithm has been tested with two different articles. One of the articles has been published in Kazakhstan and another was published in Sin Kiang Uyghur Autonomous region in China.

As a result, this algorithm was successfully able to split words into valid syllables except for Chinese and Russian words. In these two short articles a total 200 words have been used. If the number of words is increased or the type of article is changed, then the error rates may be changed and increased compared to shorted articles.

This is due to the fact that adopted words mainly appear in technical and political articles. If the following sentence is, "Men ikikünkéyin Guang Zhougha bi+rip traktor al+maq+chi+men" (I am going to Guang Zhou after two days and buy a tractor), is analyzed, the following syllables are generated.

men
ik+ki
kün
ké+yin
Guang (not solved)
Zhougha (not solve)
bi+rip
traktor (not solved)
al+maq+chi+men

Though, some words could be splitted into syllables correctly according to Uyghur syllable splitting rules, not all words can be considered Uyghur origin words. For example, following Arabic origin words can be splitted correctly with this algorithm.

Döwlet (country): döw+let
Kalem (pencil): ka+lem
Mubarek (sacred, holy): mu+ba+rek
Mektep (school): mek+tep

5. CONCLUSIONS

In this article, an approach about identifying Uyghur words according to syllabic properties is suggested and expected results have been achieved.

Although, almost all Uyghur origin words can be identified with this approach, but there are few foreign originated words also classified standard Uyghur words, as mentions in section 4. These kinds of words may be considered a special word category and have to be analyzed with a different method. With this approach not only it is possible to identify a word, it is also possible to generate random words according to the standard structure of Uyghur words.

All Turkic language words have almost the same word structure, therefore this approach may be applied to other...
Turkic Languages as well. Because of different Turkic languages having different number of characters and using different alphabet, the syllable rules maybe different relatively.

REFERENCES

1. Abdulla Tehir A, (2010). Hazirqi Zaman Uyghur Tili. Xin Jiang XelqNeshiryati, Ürümchi, China.
2. http://www.omniglot.com/writing/uyghur.htm (accessed on: 25.05.2016).
3. http://www.ukij.org (accessed on: 25.05.2016).
4. Janbaz A, Saleh I, Duval J.R (2006). An Introduction to Latin-Script Uyghur. Middle East & Central Asia Politics, Economics and Society Conference, University of Utah, USA.
5. Ofazer, K., (1995). Two-level Description of Turkish Morphology. Literary and Linguistic Computing, 9, 2, 137-148.
6. http://www.ii.metu.edu.tr/corpus, (accessed on 10 March, 2016).
7. Tantuğ A.C., Adalı E., ve Ofazer K., (2008). TürkçenedenTürkçeyeBilgisayarlıMetinÇevirisi, İTÜ Dergisi, 7, 4, 83-94.
8. Tantuğ, A.C., (2007). AkrabaveBitişken Diller ArasındaBilgisayarlıÇeviriçin Karma Bir Model. BilgisayarMühendisliğiBölümü. DoktoraTezi. İstanbulTeknikÜniversitesi, İstanbul.
9. Hamzaoğlu İ., (1993). Machine translation from Turkish to other Turkic languages and an implementation for the Azeri languages. YüksekLisansTezi. Bogazici University, İstanbul.
10. Altuntaş, K., (2000). Turkish to Crimean Tatar Machine Translation System.YüksekLisansTezi. Bilkent University, Ankara.
11. Hengirmen M (2000). TürkçeDilBilgisi, EnginYayncılık, Ankara.
12. Tömür, H., (2003). Modern Uyghur Grammar (Morphology). YıldızTeknikÜniversite, Fen-Ed Fak. T.D.E Bölümü, İstanbul, Turkiye.
13. Orhun, M., Tantuğ, A.C., and Adalı, E., (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. 811-816.
14. Orhun, M., Tantuğ, A.C., and Adalı, E., (2009c). Rule Based Analysis of the Uyghur Nouns. International Journal of Asian Language Processing, 19, 1, 33-43.