E-048

ウイグル語単音節の出現頻度とその特徴分析
Frequency of Uyghur language syllables and analysis of characteristics

アブラジャン シマイ† 猿舘 朝† 伊藤 憲三†
Abulajiang Simayi† Ashita Sarudate† Kenzo Itoh†

Abstract

Although Uyghur language is being spoken by more than ten million of people in central Asia, speech typewriter for this language has not been developed. In this paper, we analyzed syllables frequency and characteristics of Uyghur language to find the significant syllables to create an efficient database for future speech typewriter. The analyses were based on a modern Uyghur word dictionary and conversational sentences selected arbitrarily from Uyghur websites. 37,255 unique words in Uyghur modern Dictionary and 30,439 in conversational sentences were syllabified manually, resulting 2,557 and 1,437 syllables, respectively. 1,390 of these syllables share Saimaiti and Feng's (2007) which can be considered as most frequently used syllables. This number covers 98.14% of the entire Uyghur syllables. The problem arises on creating speech typewriter because of too many syllables. However, using only the 500 most frequent syllables which cover more than 90% of the entire Uyghur syllables, the speech typewriter of Uyghur language can be created efficiently.

1. Introduction

Uyghur language is classified as Altaic family sub-group of "Turkish language". It has been considered as the first language for over ten million people in central Asia including 8.5 million in northwestern China (2004 census). The alphabets of Uyghur language consist of eight vowels and 24 consonants (table 1). It uses adopted Arabic script as main writing system, called Arabic Script of Uyghur (ASU), and in other cases uses adopted Latin Script of Uyghur (LSU) as a component to ASU.

Table 1 The alphabets Uyghur language

	Arabic script of Uyghur (ASU)	Latin script of Uyghur (LSU)
Vowels	ئا،ئە،ئې،ئى،ئو،ئۇ،ئۆ،ئۈ	a, e, é, i,o, u, ö, ü
Consonants	ب، پ، ت، ج، چ، خ، د، ر، ز، ژ، س، ش، غ، ف، ق، ك، گ، ڭ، ه، ن، م، ل، ۋ، ي	b, p, t, zh, ch, x, d, r, z, j, s, sh, f, gh, q, k, g, ng, l, m, n, h, y, w

All alphabets in every Uyghur word are pronounced, so they usually correspond to phonemes. A phoneme or a group of phonemes in a word uttered concurrently constitutes a syllable. Words of Uyghur language are delimited by blanks.

Table 2 Possible structure of the syllables[1)]

Syllable structure[2)]	Example of syllable	Combination number of V and C
V	ئۇ	8
VC	ئات	192
CV	بۇ	192
CVC	كىم	4,608
VCC	ئەرز	4,608
CCV	پلا	4,608
CVV	خۇا	1,536
CVCC	خەلق	110,592
CCVC	گرام	110,592
CVVC	گۇاڭ	36,864
CCVCC	فرونت	2,654,208
	Total syllables	2,928,008

[1)] The type of syllables classified by Uyghur language textbook for middle school education (Put reference here, 2003) (C : consonant, V : vowel).
[2)] The first six syllable types are the most frequent syllables in native Uyghur words while others often occur in loanwords from Arabic, Persian, Chinese and other European languages.

†岩手県立大学大学院 ソフトウェア情報研究科
†Iwate Prefectural University, Graduated School of Software and Information Sience

Supposing a syllable would be composed of 4 or 5 alphabets at most the estimated amounts should be 2,928,008. Using general syllabification rules, Saimaiti and Feng (2007) investigated the structure of Uyghur syllable resulting 4094 unique syllables available. This result contradicts with that of Abaidula, et al. (2003) reporting 7000 syllables. Further study is needed to confirm the exact number of the syllables.

In this study, we analyze syllables frequency of Uyghur language to find the significant syllables to create an efficient database for future speech typewriter. The analyses are conducted in three steps. The first step is to collect samples of words from sources such as dictionaries and websites. The next step is to syllabify the collected samples. The final step is to compare the derived syllables to Saimaiti and Feng's (2007) and extract the significant number of syllable to create an efficient database for future speech typewriter.

2. Data

In this study, data from two sources: modern Uyghur word dictionary and conversational sentences selected arbitrarily from Uyghur websites are used for the analysis.

The modern Uyghur word dictionary contains 37,255 words and conversational sentences selected from Uyghur websites contains 30439 words.

3. Methodology

To create an efficient database for future syllable based speech typewriter of Uyghur language, syllables counted by Saimaiti and Feng (2007) needs to be validated whether it is suitable for this purpose. The validation is conducted through manual interpretation and comparison with syllables investigated from the data sources.

3.1 Validating the number of syllables counted by Saimaiti and Feng (2007)

The number of syllables counted by Saimaiti and Feng (2007) can be validated by checking the presence of recounting of same syllables, the number of miss-spelling words and errors caused by the automatic calculation algorithm conducted by Saimaiti and Feng (2007).

3.2 Syllabification rules

Each Uyghur syllable has a vowel at leas therefore every syllable in a word can be pronounced separately. The syllabification rules of Uyghur language can be sumarized as follows:

a. There is a vowel in each syllable

b. Consonants between two vowels are to be separated. If there is only one consonant between two vowels, the consonant belongs to the first vowel.

c. If there are two consonants, the front one belongs to the first syllable and the latter to the second syllable.

d. If there are three consonants, two of the front belong to first syllable and the latter to the second.

e. If there are four consonants between two vowels, two of the front belong to first syllable and the latter to the second.

f. If there are five consonants, three of the front belong to the first syllable and the rest to the second.

The above rules separate consonants between two vowels regarding to their amounts. These rules works well with original Uyghur words but not with loanwords.

3.3 Counting the most appearance syllables

To count the most appearance of syllables, words found in the data sources are cut into syllables using the syllabification rules described above. These results are then compared with Saimaiti and Feng's (2001) to produce the final most frequent used syllables in Uyghur language. This comparison is important due to the Uyghur language has a productive morph tactics yielding miss-determining of syllables boundary.

4. Results

Two native Uyghur speakers checked the original data used by Saimaiti and Feng (2007) to count the syllables. Table 3 shows the number of errors found in calculation conducted by Saimaiti and Feng (2007).

Table 3 Errors found in The number of syllables counted by Saimaiti and Feng (2007)

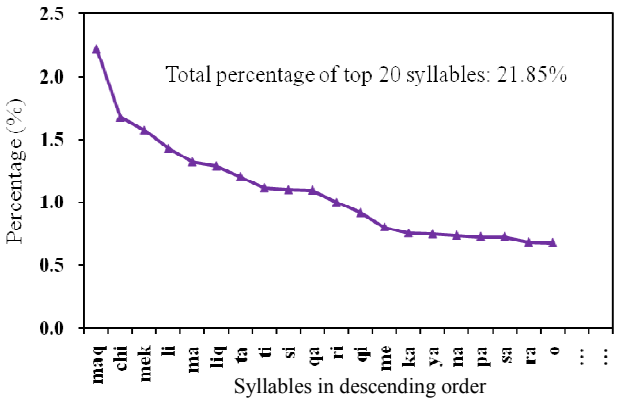
No.	Errors	Percentage
1	Recounting of same syllables	0.51%
2	Miss-spelling words	7.57%
3	Automatic algorithm	6.57%

Table 3 indicated that the possibly correct numbers of the syllables counted by Saimaiti and Feng (2007) are approximately 3,495.

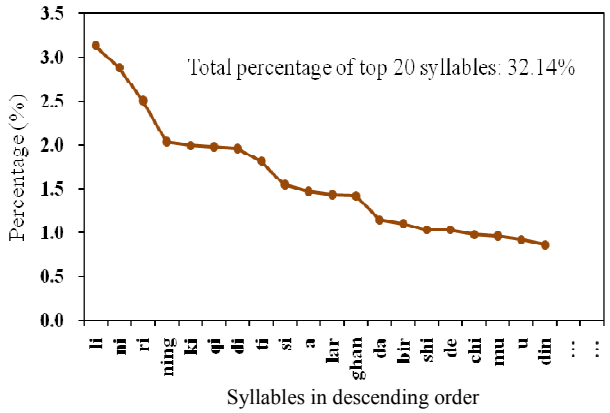
Using syllabification rules, 2,557 syllables are found from the modern Uyghur word dictionary, and 1,437 syllable from Uyghur websites. Comparing these number of syllables with Saimaiti and Feng's (2007), we found that 1,390 syllables are shared. We considered that the 1,390 syllables are the most frequent syllables for Uyghur language. Fig. 1 shows the top 20 in the number of syllables in (a) the modern Uyghur word dictionary, (b) Uyghur websites and (c) Saimaiti and Feng's

(2007), arranged with their frequency appearances in descending order. Same syllables can be observed through fig.1(a)-(c).

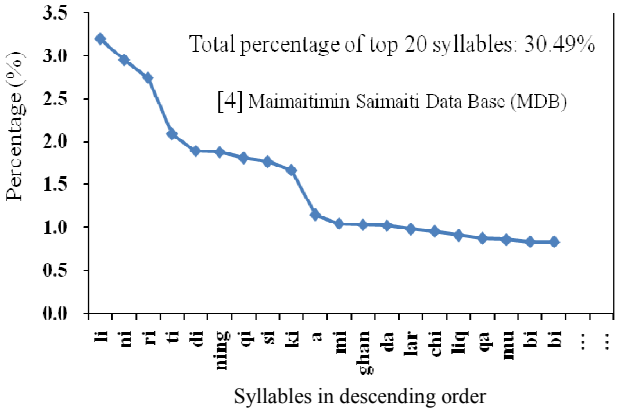
Fig.2 shows cumulative distribution of the number of syllables in the modern Uyghur word dictionary, Uyghur websites and Saimaiti and Feng's (2007) arranged with their frequency appearances in descending order. The most frequent 1,390 syllables take 98.14% of Saimaiti and Feng's (2007). This number is still relatively high to create a speech typewriter. However, taking only approximately 90% of the most frequent syllables (500 syllables) may solve this problem.



(a) The top 20 in the number of syllables in Uyghur word dictionary.



(b) The top 20 in the number of syllables in Uyghur websites.



(c) The top 20 in the number of syllables in Saimaiti and Feng's (2007).

Figure 1 The top 20 in the number of syllables in each source.

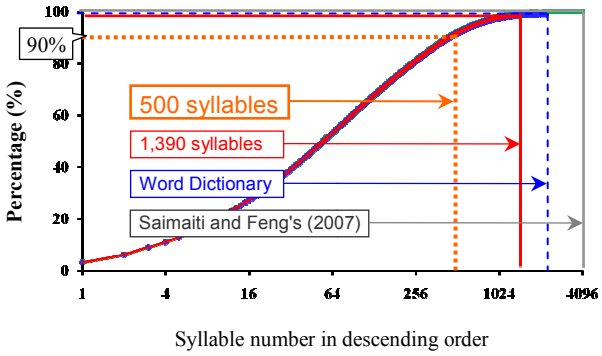


Figure 2 Cumulative distribution of the modern Uyghur word dictionary, Uyghur websites and Saimaiti and Feng's (2007)

5. Conclusion

In this research, we have introduced a method to extract the significant number of syllable to create data base for future speech typewriter. Our results provide a possibility to create efficient text data for speech typewriter by using only the 500 most frequent syllables.

References

[1] Saimaiti, M., Feng, Z., "A Syllabification Algorithm and Syllable Statistics of Written Uyghur", Proceedings of the Corpus Linguistics Conference, CL2007.

[2] Abaidula, Y., Rezwangul, Sali, A., "The Research and Development of Computer Aided Contemporary Uighur Language Tagging System", Journal of Chinese Language and Computing 15 (4), pp. 203-210, 2003.

[3] "中学校一年生のためのウイグル語の教科書", 新疆教育版, 第 4 版, 2003.

[4] 伊藤 憲三, 佐藤 大和, "会話音声における日本語音韻の出現頻度特性", 日本音響学会講演論文集, pp. 151-152, 昭和 63 年.

[5] Vaux, B., "Disharmony and derived transparency in Uyghur Vowel Harmony", Harvard University, 2000.

[6] Hoshur Islam, Rehime Tursun, "A study on Uyghur language TTS system", pp.1-8, 2005.

[7] Waris Abdukerim Janbaz, Imad Saleh, "Uyghur language processing on the Web", AIML 06 International Conference, pp.13-15. 2006.