# A Syllabification Algorithm and Syllable Statistics of Written Uyghur

Maimaitimin Saimaiti[1] and Zhiwei Feng[2]

## Abstract

In this paper, a syllabification algorithm for written Uyghur is presented, and various statistic results about syllable structure of Uyghur are analyzed based on the corpus. The algorithm, by means of an abstract computational structure, implements a set of syllabification rules, which is simple, and achieved rather high preciseness. Experiment on a random sample shows that the syllabification algorithm achieves 98.7 percent word accuracy on word tokens, 99.2 percent on word types, and 99.1 percent syllable accuracy. 30169 words in Uyghur Dictionary and 2,558,810 words from a corpus are syllabified and various aspects of syllable structure are analyzed on the base of syllable statistic results. Statistics shows that corpus based approach in computational phonology produces good result.

## 1 Introduction

Automatic syllabification, or detecting syllable boundaries in words, is an important problem to solve since it has applications in automatic speech recognition, text-to-speech systems, and corpus statistics. A great number of diverse algorithms have been proposed for syllabification in different languages and many researches has been done on these languages. However, few works has been reported on syllabification in Uyghur and its syllable structure. Uyghur is a Turkic language with about 10 million speakers mainly in the Xinjiang Uyghur Autonomous Region of China, in Central Asia, and also in other countries. It has rather complex syllable structure that has not been studied very deeply, especially using corpus. Syllabification tools are essential to manage large amount of words in a corpus.

Until now, diverse syllabification algorithms have been presented in different papers. Shankar Ananthakrishnan built a statistical syllabification algorithm that uses Supervised and Unsupervised learning (Shankar Ananthakrishnan, 2004); Karin Müller described a phonological probabilistic context-free grammar for syllable

[1] Department of Applied Linguistics, Communication University of China
*e-mail*: tilqin@yahoo.com.cn

[2] Applied Linguistics Institute, Ministry of Education of China
*e-mail*: zwfengde@hotmail.com

structure of German words (Karin Müller, 2001a, 2001b); Robert Bannert applied a rule-based approach that uses the Principle of Maximum Onset for spoken standard Swedish (Bannert R., 1998); Ouellet and Dumouchel introduced Heuristic Syllabification method (Ouellet, P., Dumouchel, P., 2001); Lorenzo Cioni described an algorithm for the syllabification of Written Italian (Lorenzo Cioni, 1997). Among other approaches that deal with syllable structure, there are example-based approaches ((Hall (1992), Wiese (1996), Féry (1995), Kenstowicz (1994), Morelli (1999)), symbolic approaches (Belz, 2000), connectionist phonotactic models (Stoianov and Nerbonne, 1998), stochastic models de-scribing partial structures (Pierrehumbert (1994), Coleman and Pierrehumbert (1997)), or application-based approaches for syllabification (Van den Bosch, 1997) or text-to-speech systems (Kiraz and Möbius, 1998), for more details see (Karin Müller. 2001a, 2001b).

Their methods in essence can be divided into two categories. One approach operates in a rule-based framework. In this case, the program parses the input phoneme sequence according to some predetermined set of rules. In this approach, knowledge about syllables structure can be built into the syllabifier. Statistical approaches, on the other hand, assume very limited knowledge of phonotactic rules and constraints. Rather, they are provided with a large corpus of data (which may or may not be labeled with the correct segmentation), and the program automatically learns the rules by estimating probability distributions over a parameter set. This learning may be supervised (when the training data is labeled) or unsupervised (when the training data is not labeled, and the program attempts to uncover patterns automatically). Statistical techniques are particularly attractive when there is a large quantity of training data, but only scant knowledge of the phonotactics of the language. In this study, rule-based algorithms are preferred for syllabification Uyghur words when there is not available training data.

In our study, we developed a syllabification algorithm for written Uyghur in order to analyze the syllable characteristics of written Uyghur. The algorithm, by means of an abstract computational structure, implements a set of syllabification rules. Experiment on a random sample shows that the syllabification algorithm achieves 98.7 percent word accuracy on word tokens, 99.2 percent on word types, and 99.1 percent syllable accuracy. After then, 30169 words in Uyghur Dictionary and 2,558,810 words from a corpus are syllabified and various aspects of syllable structure are analyzed on the base of statistic results.

The rest of the paper is organized as follows. We introduce syllable structure of Uyghur and present our syllabification method in section 2. In Section 3, the statistic results and analyses of the results are given. Finally, we conclude the paper.

## 2 Syllabification Algorithm

### 2.1 Uyghur syllables and syllabification

First, we give an overview about Uyghur language and its phonetics. Uyghur belongs to the Eastern branch of the Turkic group of the Altaic language family. 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.

Modern Uyghur has 8 vowels which is symmetrical around the axes of backness, roundness and height: /a, æ, e, i, o, u, ø, y/ which corresponds to *a, e, é, i, o, u, ö, ü* in LSU or ئا، ئە، ئې، ئى، ئو، ئۇ، ئۆ، ئۈ in ASU. Modern Uyghur has 24 consonants: /b, p, t, ʤ, ʧ, x, d, r, z, ʒ, s, ʃ, f, ʁ, q, k, g, ŋ, l, m, n, h, j, w/ which are corresponds to *b, p, t, zh, ch, x, d, r, z, j, s, sh, f, gh, q, k, g, ng, l, m, n, h, y, w* in LSU or ب، پ، ت، ج، چ، خ، د، ر، ز، ژ، س، ش، ف، غ، ق، ك، گ، ڭ، ل، م، ن، ه، ي، ۋ in ASU. The consonants have a voicing distinction. Like other Turkic languages, the Uyghur has vowel harmony, and is not a tonal language.

In grammatical aspects, Uyghur is an agglutinative language, and a number of suffixes are used for inflection and derivation. The category of aspect is expressed analytically. Uyghur is an SOV language.

Uyghur has 12 different syllable types as following: V, CVC, VC, CV, CVCC, VCC, CCVCC, CCVC, CCV, CVVC, CVCCC and CVV. First six syllable types are most frequent syllables in native (Turkic) words, others are often occurs in loanwords from Arabic, Persian, Chinese and other European languages. The stress in Uyghur words falls in general on the last syllable of the word (there are quite a few exceptions to this rule). There is at least one vowel in each syllable.

A syllable is defined as a unit of spoken language bigger than a speech sound (phoneme), and is made up of three components: a nucleus, which consists of a single vowel or syllabic consonant, optionally surrounded by one or more consonants. The consonants that precede the nucleus are collectively referred to as the onset, while those that succeed it are called the coda. The nucleus and coda are sometimes lumped together to form what is called the rhyme (Shankar, 2004). The correct syllabification of a phoneme sequence is not arbitrary, but is subject to certain phonological constraints. Especially in spoken language, the same phoneme sequence may be syllabified differently, but we ignore this detail and assume that there exists one canonical syllabification for every phoneme sequence. In polysyllabic words, the syllables are divided into syllables appearing word-initially, word-medially, and word-finally. For example, look at the syllabification of the words *ishchilar* (workers) on the left and *yardemchi* (helper) on the right. (Figure 1)

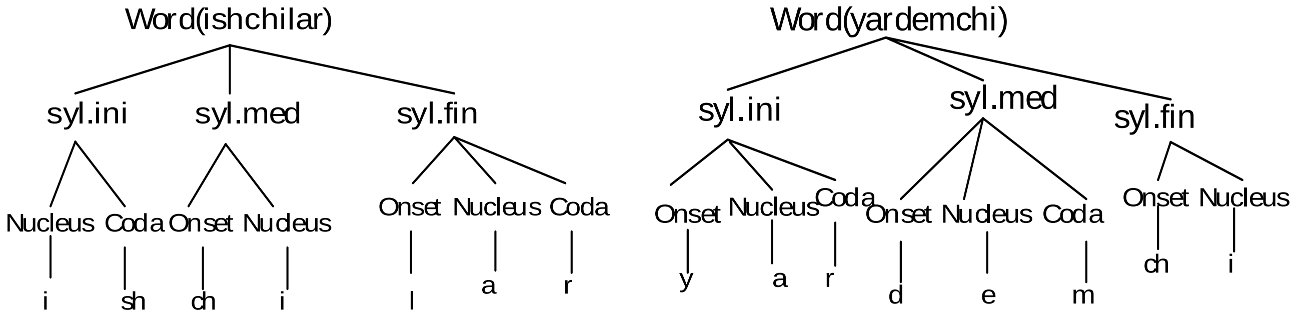

**Figure 1**: syllable structure of two Uyghur words

### 2.2 Syllabification rules

The construction of syllables in Uyghur is very regular; hence it is not difficult to find such rules intuitively or by trial. Our syllabification algorithm for written Uyghur implements a set of syllabification rules that are shortly outlined in what follows. We note that C denotes a generic consonant whereas V represents one of the eight vowels of written Uyghur. We now examine some parts of the syllabification rules:

Hamza on a tooth " ئ" is never listed separately in ASU, due to the fact that it is considered an integral part of initial form of vowels. But, in fact it does serves as the indication of syllable boundary inside a word. Presence of the letter " ئ" (a hamza on top of a tooth) inside a word is indicated through the use of an apostrophe (') in LSU.

Therefore, the first rule of syllabification is that insert a "\" (reverse solidus) as the symbol of syllable boundary if there is a hamza in ASU or an apostrophe in LSU inside a word.

The Second rule states that whenever there is a combination of the form VCCV the algorithm produces the syllabification VC\CV by inserting a reverse solidus between two consonants. The groups CVCCV, CVCCVC, CCVCCVC are syllabified as CVC\CV, CVC\CVC, CCVC\CVC separately by this rule.

Another rule states that the group VCV is syllabified as V\CV. That means that a consonant between two vowels belongs to the second syllable. The groups such as CVCV, VCVC, CVCVC, CVCVCC and CCVCV are syllabified as CV\CV, V\CVC, CV\CVC, CV\CVCC and CCV\CV separately by this rule.

Rule 4: the group VCCCV is syllabified as VCC\CV though there are exceptions such as the loan word *kompyutér* (computer) that must be syllabified as *kom\pyu\tér*. CVCCCV, VCCCVC and CVCCCVC are can be syllabified by this rule.

Rule 5: The group VCCCCV is syllabified as VCC\CCV with some exceptions

Rule 6: The group CVV or CVVC is syllabified as one syllable due to the fact that they are often loan words from Chinese, and are monosyllabic words though there are some Uyghur native words that have consequent two vowels, but the two vowels are always separated by hamze " ئ" in native words, which are syllabified based on the first Rule.

### 2.3 Syllabification process

The syllabification algorithm for written Uyghur is of deterministic type and it is based upon the use of recursion and of binary tree in order to detect the boundaries of the syllable within each word.

The algorithm is composed of input-output module that handle the input or output of words to be syllabified, and syllabification module that implements the binary tree based on the syllabification rule discussed above. The algorithm defines the path from the root to a leaf to which corresponds a syllabification rule that allows the definition of a syllable boundary.

The direction of Syllabification within each word is right-to-left, opposite to normal order. Armin Mester and Jaye Padgett (1994) attempt to explain directionality effects on syllabification result. In their work, Right-Left (R-L) syllabification and Left-Right (L-R) syllabification were compared and directional syllabification effects were systematically explored and analyzed. According to them, two kinds of syllabifications produce different result. In our study, we also considered the two possible directions, at last, adopted R-L syllabification because it is simpler and more accurate in some cases than L-R.

The algorithm scans the words from right to left and when reach the second vowel inserts the syllable boundary symbol into the correct position according to the syllabification rules. The process goes on till the last syllable is reached. This characteristic defines the algorithm as recursive.

We now inspect the syllabification process of the word *balilarning* (of children):

| Start: | Syllabified part | Rest part | Current stream | Result | Rule |
|---|---|---|---|---|---|
| 1: | 0 | balilarning | arning | ar\ning | 2 |
| 2: | \ning | balilar | ilar | i\lar | 3 |
| 3: | \lar\ning | bali | ali | a\li | 3 |
| 4: | \li\lar\ning | ba | ba | \ba | |
| End: | \ba\li\lar\ning | 0 | 0 | | |

## 2.4 Experiment and error analyses

To evaluate our syllabification algorithm, randomly selected 5000 words as a testing sample from the large corpus are syllabified automatically and the results are checked manually for errors.

The test result shows that the syllabification algorithm for written Uyghur achieves 98.7 percent word accuracy on word tokens, 99.2 percent on word types, and 99.1 percent syllable accuracy. In this report, the word accuracy indicates the fraction of words in the test that were correctly syllabified and the syllable accuracy indicates the fraction of syllables correctly identified. For obvious reason, the syllable accuracy is higher than the word accuracy. In test, Out of 5000 words, sixty-seven words contained wrongly predicted syllable boundaries.

According to our investigation on the errors, the most frequent errors occur with the loanwords due to the reason that they have the different syllable structure from the native words and their syllabification is not very regular. For example:

*Kompyutér* (computer) should be syllabified as *Kom\pyu\tér*, not as *Komp\yu\tér,* the native word *dostluqing* (your friendliness), with same structure, was syllabified as *dost\lu\qing* which is correct. *Zhinshyang*, person name from Chinese origin, was wrongly syllabified as *Zhinsh\yang*, it should be *Zhin\shyang.* A further error is found with syllable boundaries occurring in conjunction with suffixes.

*Tekstning* (of text), for instance, should be syllabified as *Tekst\ning* not as *Teks\tning* because in this word "*-ning*" is a suffix which should belong to another syllable. However, syllabification does not always mirror the morphological structure of words as the next example. *qolung* (your hand) is syllabified as *qo\lung*, which must be *qol\ung* if the suffix "*-ung*" is taken into account.

## 3 Syllable statistics of written Uyghur

In this part, first we introduce the resources on which the syllabification task was done. Second, we concentrate on a quantitative result of syllable statistic and then analyze linguistically the statistic result.

### 3.1 Corpus

To analyze the syllable structure of Written Uyghur, the syllabification algorithm was developed and corpus-based approach was adopted. In our study, we syllabified more than two million words from two sources, the first one is 30169 words in *Uyghur tilining izahliq lughiti* Contemprary Uyghur Dictionary in which only stem forms of words (or lemmas) are listed, the another one is 2,558,810 words from the Uyghur Corpus that had been constructed from 2003 to 2006 at Xinjiang University. The former one almost includes all lexemes in contemporary Uyghur language; the latter one includes texts from all subject areas of writing language such as fiction, news, science, religion, law, society, art .etc. The Uyghur Corpus has been carefully composed to ensure that all text type is represented. In our study, words from five main text types are syllabified and their frequencies are counted. The text type refers to the different levels of language that may be used in different contexts. The five text types are academic papers (15 percent), newspaper reports and opinion pieces (27 percent), corporate websites (25 percent), magazine articles (13 percent), novels and short stories (20 percent).

### 3.2 Statistic result and analyses

In this section, various statistics about the syllables that were compiled on the basis of output of the syllabification described above, in written Uyghur are given. We differentiate syllables in monosyllabic words from syllables in word initial, word medial and word final. Syllable statistics of words in corpus and words in the Dictionary are also compared in some cases to better illustrate the syllable characteristics of written Uyghur. Unfortunately, due to space constraints and the large size of statistical data, only preliminary results can be presented.

### 3.2.1 Unique syllables and their frequency

Definite number of syllables in written Uyghur is not clear due to its complexity and immature study in Uyghur phonology. In this study, trough statistics of the corpus, we tentatively found 4094 unique syllables in written Uyghur. More precisely, the 2,558,810 word tokens are consisted of 7,116,023 syllable tokens or 4094 syllable types. The numbers and frequencies of syllable types in different positions are not same. The details are as Table 1.

| Position | Mono-syllabic | Word initial | Word medial | Word final | Anyplace |
|---|---|---|---|---|---|
| type | 1371 | 2555 | 2164 | 2547 | 4094 |
| token | 294876 | 2263932 | 2263932 | 2293283 | 7116023 |
| average | 215.1 | 886.1 | 1046.2 | 900.4 | 1738.2 |

**Table 1**: Number of syllable types in different positions

From the Table 1, we can see that syllable types in monosyllabic words are 1371 that is less than initial and final syllables which means that the syllables in word-initial and word-final are more changeable. The table explains the diversity of syllables in different position. The 4094 unique syllables are obtained in our study, however their frequencies are various from one syllable to another. Therefore, 4094 syllables were divided into seven groups (Table 2) based on their frequencies in order to better understand.

| Group name | Frequency | Accumulative frequency | Accumulative percentage | Number |
|---|---|---|---|---|
| 1 | 1-20(≥58060) | 2168515 | 30.47 | 20 |
| 2 | 21-100(≥14944) | 4447972 | 62.51 | 80 |
| 3 | 101-500(≥1769) | 6517888 | 91.59 | 400 |
| 4 | 501-672(≥1000) | 6748703 | 94.84 | 172 |
| 5 | 673-1547(≥100) | 7077477 | 99.46 | 875 |
| 6 | 1548-2461(≥10) | 7110429 | 99.92 | 914 |
| 7 | 2462-4094(<10) | 5594 | 0.08 | 1633 |
|  | Sum | 7116023 | 100.00 | 4094 |

**Table 2**: syllable frequency and grouping

Second column in table 2 includes number of syllables in each group and their frequencies in bracket, the third and forth column are accumulative frequency and percentage respectively, the last one is number of syllable types in each group.

Among 4094 syllables, 2461 syllables appeared more than ten times which covers the 99.92 percent of all words in corpus. 1547 syllables appeared over 100 times and 672 syllables appeared over 1000 times. 500 syllable types that appeared over 1769 are account for more than 90 percent of syllables. Sixty-two percent words are consisted of only 100 syllables; the twenty most frequent syllables are account for the 30.47 percent of syllable tokens. The twenty most frequent syllables and their frequencies are as shown in the Figure 2. Note that most of them are of C+i syllable structure.

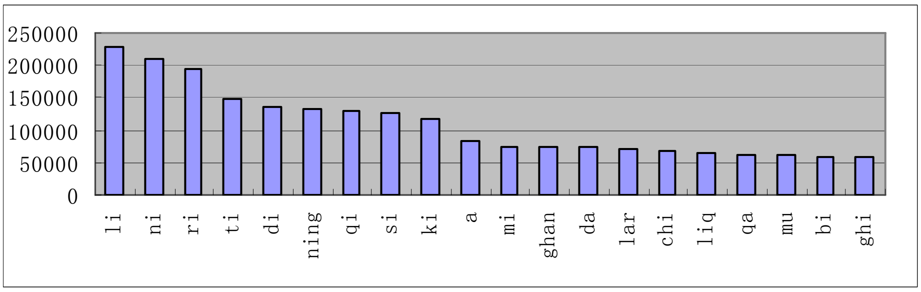

**Figure 2**: twenty most frequent syllables in Uyghur.

### 3.2.2 Word length

Number of syllables in each word is also counted to analyze the word length in Uyghur. In Figure 3, three kinds of statistic results are shown, namely word types, word tokens and lexemes in the Dictionary.

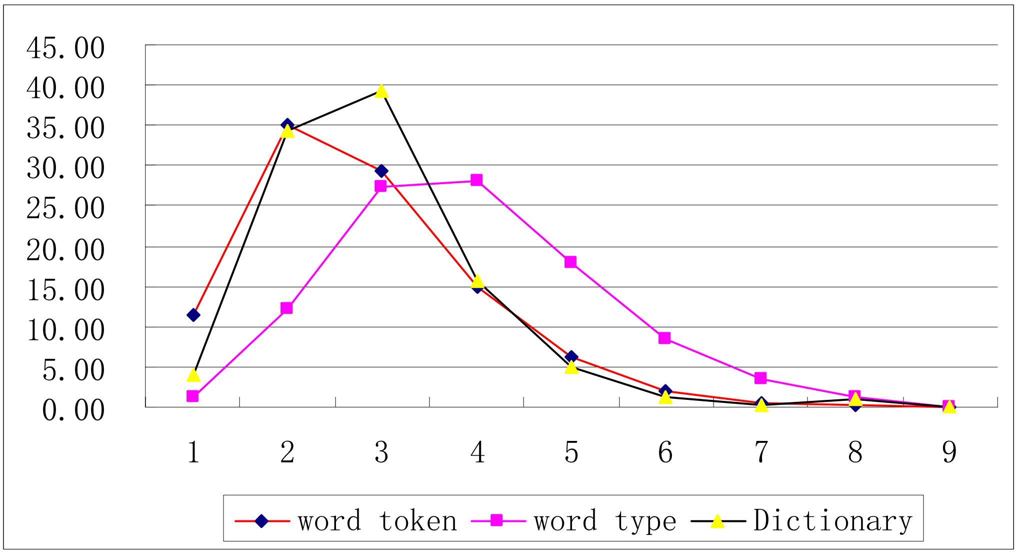

**Figure 3**: Word length by number of syllables

According to statistic result, longest word in Uyghur is consisted of nine syllables; average length of the word is 2.87 (7116023/2558810) syllables. From the figure, it is not difficult to find that the most frequent words are bi-syllabic words (35.2%) though their number (13418, 12.2%) is less than that of tri-syllabic words (29978, 27.3%) and quadro-syllabic words (30809, 28%). The simpler words with fewer syllables tend to be more frequent. However, tri-syllabic words in dictionary are more than bi-syllabic word and quadro-syllabic words. Another obvious result of the statistics is that bi-syllabic words (34.2%) and tri-syllabic words (39.4%) are more than all other polysyllabic words in stem forms; however, quadro-syllabic words (28%) and tri-syllabic words (27.3%) are account for the larger proportion of the word types. Percentage of word types with more than three syllables is higher than that of stem words with same syllable number. This indicates that most of Uyghur words appear in inflected forms in the sentences.

### 3.2.3 Syllable complexity

As shown Table 3 and Table 4, Uyghur has rather complex syllable structure although some syllable structures are not frequent. According to statistic result, most frequent syllables in written Uyghur are CVC (53.3%) and CV (38.4%) structures in stem form of words while CV (50.01%) and CVC (40.44%) structures are most frequent in inflected words. The reason for this phenomenon is that some words change their syllable structure after inflectional suffixes are added. For example, *qol* (hand) that is of CVC structure becomes *qoli* (his hand) by the suffixation of *–i*, which has the structure of CV+CV. VC, V and CVCC syllable structures are less frequent, their frequencies are 3.74%, 2.98% and 1.36% in stem form, or 4.5%, 4.2% and 0.6% in inflectional form respectively.

| Structure | Mono-syllabic | Word initial | Word medial | Word final | Sum | Percent age |
|---|---|---|---|---|---|---|
| CV | 40 | 12706 | 13408 | 7308 | 33462 | *38.4* |
| CVC | 867 | 10867 | 13662 | 21007 | 46403 | *53.3* |
| VC | 78 | 2712 | 291 | 177 | 3258 | *3.74* |
| V | 2 | 2187 | 371 | 33 | 2593 | *2.98* |
| CVCC | 215 | 366 | 223 | 377 | 1181 | *1.36* |
| VCC | 27 | 61 | 4 | 10 | 102 | *0.12* |
| VCCC | 0 | 14 | 0 | 0 | 14 | *0.02* |
| CVCCC | 5 | 11 | 3 | 3 | 22 | *0.03* |
| CVVC | 2 | 6 | 2 | 13 | 23 | *0.03* |
| CVV | 0 | 1 | 0 | 3 | 4 | *<0.01* |
| CCVCC | 0 | 1 | 0 | 0 | 1 | *<0.01* |
| CCVC | 1 | 1 | 0 | 0 | 2 | *<0.01* |
| Sum | 1237 | 28933 | 27964 | 28931 | 87065 | |

**Table 3**: Distribution of different syllable structures in the Dictionary

| Structure | Mono-syllabic | Word initial | Word medial | Word final | Sum | Percent-age |
|---|---|---|---|---|---|---|
| CV | 55437 | 1026305 | 1487130 | 989660 | 3558532 | *50.01* |
| CVC | 186834 | 683921 | 753262 | 1253800 | 2877817 | *40.44* |
| VC | 26376 | 268211 | 15363 | 11329 | 321279 | *4.515* |
| V | 7540 | 259603 | 31393 | 0 | 298536 | *4.195* |
| CVCC | 14963 | 18749 | 5120 | 6519 | 45351 | *0.637* |
| VCC | 928 | 3523 | 181 | 416 | 5048 | *0.071* |
| CVVC | 2652 | 2674 | 548 | 1116 | 6990 | *0.098* |
| CCVC | 44 | 220 | 0 | 0 | 264 | *0.004* |
| CVV | 30 | 198 | 182 | 755 | 1165 | *0.016* |
| VCCC | 7 | 177 | 4 | 0 | 188 | *0.003* |
| CVCCC | 63 | 175 | 78 | 333 | 649 | *0.009* |
| CCV | 0 | 153 | 0 | 0 | 153 | *0.002* |
| CVVCC | 0 | 13 | 6 | 4 | 23 | *<0.001* |
| CCVCC | 2 | 10 | 0 | 0 | 12 | *<0.001* |
| CVCCCC | 0 | 0 | 16 | 0 | 16 | *<0.001* |
| Sum | 294876 | 2263932 | 2293283 | 2263932 | 7116023 | |

**Table 4**: Distribution of different syllable structures in the Corpus

Other structures such as CCVC, CCVCC, CVVC and CVV have very law frequency because most of them are unique to foreign words. From the tables, we also can see that same syllable structures have very different frequencies in different position and the simpler syllable structures are more likely to be more frequent, even some syllable structures are occurs in specific position. For instance, CCVC structures are always found in monosyllabic words or word initial.

### 3.2.4 Nuclei complexity

As we discussed above, there is one or two vowels in each syllable, more than two vowels within one syllable is not permitted in Uyghur phonology. Syllable types with one vowel are most common, syllable types with two vowels are only appears in foreign words which only accounts for less than 1% of all syllables. Therefore, most of nuclei are rather simple; statistics of nuclei are given in Table 5.

| Nuclei | initial | Percentage | final | Percentage | medial | Percentage | mono | Percentage | total | Percentage |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 540812 | *23.86* | 504217 | *22.25* | 435838 | *19.00* | 46670 | *15.68* | 1527537 | *21.44* |
| E | 429365 | *18.94* | 411125 | *18.14* | 249523 | *10.88* | 83742 | *28.14* | 1173755 | *16.48* |
| I | 431878 | *19.05* | 1049398 | *46.31* | 1253174 | *54.63* | 65443 | *21.99* | 2799893 | *39.30* |
| *É* | 195713 | *8.63* | 3916 | *0.17* | 29371 | *1.28* | 4018 | *1.35* | 233018 | *3.27* |
| o | 248536 | *10.96* | 42408 | *1.87* | 67832 | *2.96* | 26433 | *8.88* | 385209 | *5.41* |
| u | 233330 | *10.29* | 206232 | *9.10* | 185076 | *8.07* | 45292 | *15.22* | 669930 | *9.40* |
| *ö* | 105107 | *4.64* | 868 | *0.04* | 2908 | *0.13* | 15533 | *5.22* | 124416 | *1.75* |
| *ü* | 82072 | *3.62* | 47637 | *2.10* | 70301 | *3.06* | 10427 | *3.50* | 210437 | *2.95* |

**Table 5**: Statistics of nucleus in Uyghur

The most likely nuclei are in initial syllables [a, i, e] (23.86%, 19.05%, 18.94%), in medial syllables [i, a, e, u] (54.63%, 19%, 10.88%, 8.07%), in final syllables [i, a, e, u] (46.31%, 22.25%, 18.14%, 9.10%), in monosyllabic words [e, i, a, u] (28.14%, 21.99%, 15.68%, 15.22%), and in total [i, a, e] (39.3%, 21.4%, 16.5%). The less likely nuclei are [ü, ö] (3.62%, 4.64%) in initial syllables, [ö, é] (0.04%, 0.17%) in final syllable, [ö, é] (0.13%, 1.28%) in medial syllables, [é, ü] (1.35%, 3.5%) in monosyllabic words, and [ö, ü, é] (1.75%, 2.95%, 3.27%) in total.

### 3.2.5 Onset and coda complexity

Many syllables in Uyghur prefer simple onset and coda. For onset, a single consonant is found (93%), two consonants (less than 0.01%), and three consonants are not found. For codas, one consonant is observed (46%), two consonants (0.7%), and three consonants (<0.01%). Table 5 displays the onsets and codas consisting of one consonant.

**Onset statistics:** According to the statistics result, the most probable consonants in monosyllabic words are [b, y, m, x] (30.6%, 8.4%, 7.6%, 6.4%), in initial syllables [b, t, q, m] (13.7%, 12.2%, 11.8%, 9.1%), in medial syllables [l, r, t, m] (18.7%, 10.4%, 9.6%, 6.5%), in final syllables [l, n, d, r] (15.6%, 14.5%, 12.4%, 8.2%), and in total [l, t, d, n] (12.1%, 9.2%, 7.7%, 7.6%). [ng] does not appear in monosyllabic words and word initial syllables as onset, and [j, f, ng] (0.01%, 0.1%, 0.29%) are less likely to be an onset.

**Coda statistics:** in initial position, the most likely consonants are [l, r, n, y] (16.9%, 13.1%, 8.5%, 7.7%), in medial syllables [r, n, t, l] (20.7%, 17.7%, 10.6%, 8.4%), in final syllables [n, ng, r, p, q] (22.2%, 12.4%, 10.4%, 9%, 8.9%), and in monosyllabic words [r, z, ng, l] (19,3%, 10.2%, 9.4%, 9.3%). Less likely consonant are [j, f, g, zh] (0.03%, 0.05%, 0.14%, 0.23%).

| Onsets | Mono | initial | medial | final | total% |
|---|---|---|---|---|---|
| ب (b) | 30.61 | 13.72 | 2.13 | 1.12 | 6.01 |
| پ (p) | 2.17 | 3.00 | 1.56 | 0.72 | 1.68 |
| ت (t) | 5.93 | 12.23 | 9.63 | 6.77 | 9.19 |
| ج (zh) | 1.51 | 1.71 | 1.17 | 1.24 | 1.35 |
| چ (ch) | 2.88 | 2.43 | 2.40 | 2.70 | 2.53 |
| خ (x) | 6.44 | 3.78 | 0.93 | 0.31 | 1.70 |
| د (d) | 4.55 | 3.88 | 6.37 | 12.35 | 7.71 |
| ر (r) | 0.75 | 2.60 | 10.41 | 8.18 | 7.16 |
| ز (z) | 1.29 | 0.91 | 2.57 | 1.55 | 1.72 |
| ژ (j) | 0.00 | 0.04 | 0.01 | 0.00 | 0.01 |
| س (s) | 5.50 | 5.10 | 5.62 | 4.24 | 5.00 |
| ش (sh) | 4.14 | 3.41 | 2.56 | 2.44 | 2.81 |
| غ (gh) | 0.58 | 0.30 | 2.38 | 6.20 | 3.08 |
| ف (f) | 0.09 | 0.14 | 0.07 | 0.11 | 0.10 |
| ق (q) | 3.10 | 11.81 | 5.28 | 3.80 | 6.42 |
| ك (k) | 4.56 | 8.35 | 3.59 | 5.30 | 5.49 |
| گ (g) | 1.07 | 0.80 | 2.14 | 3.27 | 2.13 |
| ل (l) | 0.52 | 0.70 | 18.66 | 15.60 | 12.08 |
| م (m) | 7.56 | 9.05 | 6.51 | 4.39 | 6.50 |
| ن (n) | 2.11 | 2.99 | 4.87 | 14.50 | 7.60 |
| ي (y) | 8.40 | 7.86 | 5.09 | 2.74 | 5.15 |
| ڭ (ng) | 0.00 | 0.00 | 0.30 | 0.54 | 0.29 |
| ھ (h) | 5.52 | 4.16 | 1.81 | 0.71 | 2.20 |
| ۋ (w) | 0.72 | 1.03 | 3.94 | 1.21 | 2.09 |

| Codas | Mono | initial | medial | final | total% |
|---|---|---|---|---|---|
| ب (b) | 0.26 | 0.94 | 0.92 | 0.28 | 0.63 |
| پ (p) | 7.94 | 3.75 | 4.59 | 8.99 | 6.29 |
| ت (t) | 7.55 | 5.77 | 10.63 | 6.36 | 7.28 |
| ج (zh) | 0.37 | 0.50 | 0.14 | 0.05 | 0.23 |
| چ (ch) | 2.03 | 1.22 | 0.34 | 0.18 | 0.66 |
| خ (x) | 0.10 | 3.80 | 0.28 | 0.05 | 1.23 |
| د (d) | 0.23 | 1.48 | 0.42 | 0.37 | 0.71 |
| ر (r) | 19.32 | 13.09 | 20.65 | 10.41 | 14.28 |
| ز (z) | 10.17 | 4.01 | 2.99 | 1.97 | 3.41 |
| ژ (j) | 0.00 | 0.00 | 0.10 | 0.00 | 0.03 |
| س (s) | 1.83 | 5.01 | 2.56 | 0.92 | 2.60 |
| ش (sh) | 8.16 | 6.64 | 7.76 | 7.14 | 7.21 |
| غ (gh) | 0.38 | 1.75 | 0.22 | 1.45 | 1.17 |
| ف (f) | 0.01 | 0.06 | 0.02 | 0.05 | 0.05 |
| ق (q) | 5.74 | 4.47 | 5.38 | 8.92 | 6.52 |
| ك (k) | 1.51 | 2.89 | 2.77 | 5.32 | 3.71 |
| گ (g) | 0.03 | 0.33 | 0.13 | 0.02 | 0.14 |
| ل (l) | 9.32 | 16.89 | 8.43 | 2.60 | 8.75 |
| م (m) | 2.59 | 4.29 | 6.65 | 6.53 | 5.60 |
| ن (n) | 9.23 | 8.50 | 17.68 | 22.21 | 16.09 |
| ي (y) | 3.41 | 7.67 | 4.39 | 3.55 | 4.98 |
| ڭ (ng) | 9.40 | 3.48 | 2.36 | 12.42 | 7.13 |
| ھ (h) | 0.29 | 2.54 | 0.42 | 0.15 | 0.94 |
| ۋ (w) | 0.12 | 0.93 | 0.18 | 0.04 | 0.35 |

**Table 6**: statistics of onsets (left) and codas (right) in Uyghur

## 4 Conclusion and future work

In this paper, a syllabification algorithm for written Uyghur was presented, and various statistics result about syllable structure of Uyghur was analyzed based on the corpus. Our syllabification algorithm works on the basis of syllabification rule, which is simple and achieved rather high preciseness on syllabification. Statistics shows that corpus based approach in phonology produces good result. In future, we would like to improve the algorithm and conduct more experiments about syllable structure of Uyghur spoken language.

## Acknowledgements

We would like to thank Professor Hou Min for her advice and support in writing this paper. We would also like to Graduate School and center for corpus research on media spoken language at Communication University of China for its financial support.

## References

Armin Mester and Jaye Padgett. (1994) ‘Directional Syllabification in Generalized Alignment’. *Phonology*, Vol. 3, pp. 79–85

Bannert R. (1997) Principles of syllabification for Swedish: a methodological study. Reports from the Department of Phonetics, Umeå University. PHONUM 4, pp. 73–76.

Bannert R. (1998) Two thousand and one syllables in spoken Standard wedish: aspects of syllabification. Reports from the Department of Phonetics, Umeå University.

Jean Rahman Duval, Waris Abdukerim Janbaz. (2006) An Introduction to Latin-Script Uyghur. *Middle East & Central Asia Politics, Economics, and Society Conference*. Sept 7–9, 2006, Salt Lake City, USA.

Seung-Shik Kang and Yung Taek Kim (1994) *Syllable-based model for the Korean Morphology*, Proceedings of the 15-th International Conference on Computational Linguistics(COLING-94), vol.1, pp. 221–26. Kyoto, Japan

Karin Müller. (2001a) Automatic Detection of Syllable Boundaries Combining the Advantages of Treebank and Bracketed Corpora Training. In Proc. of ACL 2001.

Karin Müller. 2001b. Probabilistic Context-Free Grammars for Syllabification and Grapheme-to-Phoneme Conversion. In Proc. of EMNLP, Pittsburgh, PA.

Karin Müller. (2002) Probabilistic Context-Free Grammars for Phonology, Morphological and Phonological Learning. *Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, Philadelphia, July 2002, pp. 70–80.

Lorenzo Cioni. (1997) an algorithm for the syllabification of written Italian. *"V Simposio Internacional de Comunicaci.n Social", Santiago de Cuba, 22–24, 1997.*

Ouellet, P., Dumouchel, P. (2001) Heuristic Syllabification and Statistical Syllable-Based Modeling for Speech-Input Topic Identification, *Workshop on Grammar and NLP.* Montreal, Quebec, Canada, October 13–14, 2001

Reé MacKinney-Romero, John Goddardn. (2006) Syllabification Using Decision Trees, Early Results on Three Languages. Available on-line fromhttp://ccc.inaoep.mx/~tec_lenguaje06/ (accessed: 1 February 2008)

Saimaiti Maimaitimin. (2004) Study on Phonemic Combination and Syllabic Structure of Modern Uyghur. *The Journal of Xinjiang University*, Vol.4, 2004.

Shankar Ananthakrishnan. (2004) Statistical Syllabification of English Phoneme Sequences using Supervised and Unsupervised Algorithms, CS562 Term Project Report.

Shim, K. S. (1996) Automated Word-Segmentation for Korean using Mutual Information of Syllables, *Journal of KISS: Software and Applications*, pp. 991–1000

Yusup Abaidula, Rezwangul, Abdiryim Sali. (2005) The Research and Development of Computer Aided Contemporary Uighur Language Tagging System. *Journal of Chinese Language and Computing,* vol. 15 (4), pp. 203–210.