Formation of Standard Tibetan Syllables and Comparison as well as Analysis of the Statistical Results

Weilan Wang
Institute of Information Technology
Northwest University for Nationalities
Lanzhou, China
Wangweilan@xbmu.edu.cn

Duola
Institute of Information Technology
Northwest University for Nationalities
Lanzhou, China
duola@xbmu.edu.cn

***Abstract*—Considering how best it can be designed and realized through computer, the combination of all types of letters as well as its prefix, root, superscript, subscript, vowel, suffix, and farther suffix that are contained in a syllable will be transcribed by Latin after carefully identifying the attributes of Tibetan word in the statistical form. All these component parts of a syllable keep to the rules that have to be followed in the process of combining all types of letters, thereby theoretically producing normative syllables that consist of one, two, three and four letters. Normative syllables that are made up of respectively one, two, three and four characters can be found in the following numbers separately 445, 4985, 7212, 2250 and these syllables amount to 14,982 when all put together. Statistical results of Tibetan normative syllables in language databank with a size of 50 megabyte appear in these following numbers respectively: 415 single-character syllables, 2475 double-character syllables, 2423 triple-character syllables, 524 quadruple-character syllables and in total there are 5837 syllables. The findings in the experiments indicate that these syllables will turn up inconsistently across various language databases, but the most frequently occurring syllables are stably distributed while the non-frequent ones will differ in their level of presence, according to the size of database being referred to; and there is slight change in the frequency of the medium syllables. And statistical results that come up from the experiments seem to be contradictory to the number of theoretically normative Tibetan syllables available in Tibetan language, syllables in actual use only account for 39.2% of theoretically normative syllables. And syllables present in more than 90% of the texts only account for 12% of the syllables in actual use.**

***Keywords*—Tibetan, normative syllable, produce, statistical analysis**

I. INTRODUCTION

Tibetan syllable is a language unit with its root letter in the centre. Each upright unit (a consonant alone or a combination of stack) in a syllable is named character. Speaking from the perspective of composition of a syllable, there are single-character, double-character, triple-character and quadruple-character syllables in Tibetan language. What are the theoretically normative syllables like and how many are there in the Tibetan language? This is a fundamental question constantly raised about the Tibetan language in the context of research concerning Tibetan information technology. Moreover, statistical analysis of the corpus is the only available way of obtaining accurate as well as complete knowledge about the Tibetan language. Normative syllables will be of great value for the future work on the Tibetan Information Processing (TIP) to resolve issues that might arise subsequently. Our former work [1,2] did a careful statistical calculation and analysis in detail, it relate to frequency, accumulatively frequency, information entropy and relation of frequency-rank etc., the obtained high frequency data are exact and reasonable. On this basis, we further improve research methods, reduce or eliminate manual interference, increase the coverage of corpus to arrive at a more complete and accurate statistical results in the experiments. In the process of study when we actually increase corpus, no big change will take place in the high frequency syllables; The syllables of medium frequency will have necessarily change and syllables of low frequency will increase quantitatively. On one hand, it demonstrates that when the size of the corpus reaches certain quantity, the high frequent syllables will stop reacting to the continual expansion of database and the frequency of these syllables will stabilize gradually; On the other hand, it indicates that the corpus in use now is far from being able to satisfy the demands of statistical study. Corpus construction is a systematic project that can not be fully completed with our current ability over a short course of time. This paper is to focus:

(1) according as the Latin Transcribing of root letter, superscript letter, subscript letter and vowel by the stacked layer numbers of character in a syllable, the attributes of character in data structure form are described, and using the combination rules of root letter with prefix, superscript , subscript, suffix, second suffix and vowel, the normative syllables that consist of single-character syllables, double-character syllables, triple-character syllables and quadruple-character syllables are described as well as.

(2). on (1) basis, theoretically consistent normative single-character syllables, double-character syllables, triple-character syllables and quadruple-character syllables are designed and generated.

(3). statistics and analysis of frequency distribution for normative syllables in 50M corpus.

II. DESCRIPTION OF STRUCTURE OF A CHARACTER IN A TIBETAN SYLLABLE

There are many ways available to describe a normative Tibetan syllable, for instance, description of a syllable in terms of seven different component parts of its own structure. Due to the fact that a syllable is made up of its character put together, the method employed in this paper in the creation of a syllable is to firstly design structural attributes of character, with which we will be able to fully describe all characters. We have to rely on the attribute value of a character to describe a syllable. Descriptions of structural attributes of character are as follows:

```
typedef struct _Character
{
  TCHAR sTibet[3];//it denote Tibetan character,or character
  WORD sWord;  //it denote ISN of character, or ISN
  TCHAR sRoot[5];  //it denote Latin Transcribing of root
                   letter, or root
  TCHAR sSuper;  //it denote Latin Transcribing of
                   superscript letter, or superscript
  TCHAR sSub;  // it denote Latin Transcribing of subscript
                   letter, or subscript
  TCHAR sVowel; // it denote Latin Transcribing of vowel
                   letter, or vowel
  int iNumber;  // it denote the number of layer a character, or
                   number of layer
}Character;
```

The attribute list of all the characters can be generated by this structure, so that we express standard syllable with certain structure.

III. DESCRIPTION OF STRUCTURAL COMBINATION OF A TIBETAN SYLLABLE

Table 1 shows the commonly used Tibetan letters and corresponding Latin coding, ཨ should be represented with A when it functions as a long vowel. Descriptions are made accordingly based on the attributes of the character and rules involved in the formation of syllables of different length. Single-character syllable needs to be described in the number of layer, superscript, subscript and vowel. Each of the Latin coding for root letters below has 'a' together with them, and other component parts such as prefix, superscript , subscript, suffix and farther suffix do not go together with 'a'. These rules that have to be strictly followed in the process of combining all types of letters together to form a syllable are proposed in reference to many both classic and typical works devoted to Tibetan grammar , for example, documents [3][4][5][6][7]. But traditional books about Tibetan grammar are a little bit different from one another in the way how they put rules involved in the formation of a syllable. Based on these ideas mentioned above, this paper is going to do further research and more detailed study according to the practical needs.

Table 1. Coding of 30 letters and 4 vowels for Tibetan

ཀ	ཁ	ག	ང	ཅ	ཆ	ཇ	ཉ	ཏ	ཐ	ད	ན	པ	ཕ	བ	མ	ཙ
ka	kha	ga	nga	ca	cha	ja	nya	ta	tha	da	na	pa	Pha	ba	ma	tsa
ཚ	ཛ	ཝ	ཞ	ཟ	ཡ	ཡ	ར	ལ	ཤ	ས	ཧ	ཨ	ི	ུ	ེ	ོ
tsha	dza	wa	zha	za	va	ya	ra	la	sha	sa	ha	a	i	u	e	o

A. Normative single-character syllables

Single-character syllable can be divided into four types as following:

1. Single-layer character: thirty consonant letters
2. Two-layer character

(1) Each of 30 consonant letters separately combined with one of four vowels

(2) The form of "root letter + subscript letter":

（ｉ）root letter are " *ka, kha, ga, pa, pha, ba, ma*" and subscript is "*y*"；

（ii）root letter are "*ka, kha, ga, ta, tha, da, na, pa, pha, ba, ma, sha ,sa ,ha*" and subscript is "*r*"；

（iii）root letter are "*ka, ga, ba, ra, sa, za*" and subscript is "*l*"；

（iv）root letter are "ka, kha, ga, ca, nya, ta, da, tsha, zha, za, ra, la, sha, sa, ha" and subscript is "*w*".

（3）The form of "root letter + superscript letter":

（ｉ）root letter are "ka, ga, nga, ja, nya, ta, da, na, ba, ma, tsa, dza" and superscript is "*r*"；

（ii）root letter are "ka, ga, nga, ca, ja, ta, da, pa, ba, ha" and superscript is "*l*"；

（iii）root letter are "ka, ga, nga, nya, ta, da, na, pa, ba, ma, tsa" and superscript is "*s*".

（4）The form of "root letter + long vowel": root letter are "ca, cha, zha" and long vowel "A".

3. Three-layer character

（1）the form of "root +superscript + subscript "

（ｉ）the root is one of "ka, ga, ma", superscript is "r" and subscript is "y";

（ii）the root is one of "*ka, ga, pa, ba, ma*", superscript is "*s*" and subscript is "*y*";

（iii）the root is "*ka, ga, pa, ba, ma*" superscript is "*s*" and subscript is "*r*";

（iv）the root is "*tsa*", superscript is "*r*" and subscript is "*w*".

（2）the form of "root +superscript +vowel ":

Vowel is "i":

（ｉ）the root is one of "*nya, ta, da, ma, tsa, dza, ja*" and superscript is "*r*";

（ii）the root is one of "*ca, ja, ta, da*" superscript is "*l*";

（iii）the root is one of "*ga, nya, ta, da, ma*" superscript is "*s*".

Vowel is "u":

（ｉ）the root is one of "*ka, ga, nga, ja, ta, da, na, ma, tsa, dza*" and superscript is "*r*";

（ii）the root is one of "*ka, ta, da, ba, ha*" and superscript is "*l*";

（iii）the root is one of "*ka, ga, nga, nya, ta, da, na, pa, ba, ma*" and superscript is "*s*".

Vowel is "e":

（ｉ）the root is one of "*ka, ga, ja, nya, ta, ma, tsa, dza*" and superscript is "*r*";

（ii）the root is one of "*ca, ta, da, ha*" and superscript is "*l*";

（iii）the root ia one of "*ka, ga, nya, ta, da, na, pa, ba, ma*" and superscript is "*s*".

Vowel is "o":

（ i ）the root is one of "*ka, ga, nga, ja, nya, ta, da, na, ba, ma, tsa, dza*" and superscript is "*r*";

（ ii ）the root is one of "*ka, ca, ja, ta, da, ha*" and superscript is "*l*";

（iii）the root is one of "*ka, ga, nga, nya, ta, da, na, pa, ba, ma, tsa*" superscript is "*s*".

（3）the form of "root + subscript +vowel":

Vowel is "i":

（ i ）the root is one of "*ka, kha, ga, pha, ba, ma*" and subscript is "*y*";

（ ii ）the root is one of "*ka, kha, ga, tha, ta, da, pa, pha, ba, ma, sa, ha*" and subscript is "*r*";

（iii）the root is "*ga*" or "*ra*" and subscript is "*l*".

Vowel is "u":

（ i ）the root is one of "*ka, kha, ga, pha, ba, ma*" and subscript is "*y*";

（ ii ）the root is " *ka, kha, ga, da, pha, ba, sa, ha*" and subscript is "*r*";

（iii）the root is "*ka, ga, ba, za, ra, sa*" and subscript is "*l*".

Vowel is "e":

（ i ）the root is one of "*ka, kha, ga, ba, pha, ma*" and subscript is "*y*";

（ ii ）the root is one of "*ka, kha, ga, pha, ta, da, ba*" and subscript is "*r*";

（iii）the root is one of "*ka, ga, sa*" subscript is "*l*".

Vowel is "o":

（ i ）the root is one of "*ka, kha, ga, pa, pha, ba, ma*" subscript is "*y*";

（ ii ）the root is one of "*ka, kha, ga, ta, da, pa, pha, ba, sa*" subscript is "*r*";

（iii）the root is one of "*ka, ga, ba, za, ra, sa*" and subscript is "*l*".

（4）the root is "*ga*", subscript is "*r*" and farther subscript is "*w*".

4．Four-layer character

The form of "root +superscript + subscript +vowel":

Vowel is "i":

（ i ）the root is "*ma*", superscript is "*r*" and subscript is "*y*";

（ ii ）the root is one of "*ka, ga, pa, ba, ma*", superscript is "*s*" and subscript is "*y*";

（iii）the root is one of "*ka, ga, pa, ba, ma*" ,superscript is "*s*" and subscript is "*r*".

Vowel is "u":

（ i ）the root is "ga", superscript is "r" and subscript is "y";

（ ii ）the root is one of "*ka, ga, pa, ma*", superscript is "*s* " and subscript is "*y*";

（iii）the root is one of "ka, ga, pa, ba, na", superscript is "s" and subscript is "r".

Vowel is "e":

（ i ）the root is "*ka*", superscript is "*r*", subscript is "*y*";

（ ii ）the root is "*ka*" or "*ga*", superscript is "*s*" and subscript is "*y*";

（iii）the root is one of "*ga, na, pa, ba, ma*" ,superscript is "*s*" and subscript is "*r*".

Vowel is "o":

（ i ）the root is "*ka*" or "*ga*" ,superscript is "*r*" and subscript is "*y*";

（ ii ）the root is one of "*ka, ga, pa, ba, ma*", superscript is "*s*" and subscript is "*y*";

（iii）the root is one of "*ka, ga, pa, ma, na*" superscript is "*s*" and subscript is "*r*"..

B. Normative double-character syllables

There are 4 types for normative double-character syllables as following:

1. The number of layers of two characters is 1

First one is one of 30 consonants, second one is one of 9 suffixes except suffix "v"; especially after "la" do not follow "l".

2. First one is one of 5 prefixes and second character is 2 layers.

（1）prefix is "b" :

（ i ）the root is "*ta*" and superscript is "*l*";

（ ii ）the root is "ka" and subscript is "r";

（iii）the root is "ra" and subscript is "l";

（iv）the root is "da" and superscript is "r".

（2）prefix is "v" :

（ i ）the root is "pha" and subscript is "y";

（ ii ）the root is "da" or "pha" and subscript is "r".

（3）prefix is "g", second character is "root + vowel "and root is one of "ca, ta, da, sa, zha, za, ya, sha";

（4）prefix is "d"and root is "ga" and subscriot is "r".

（5）prefix is "m"and second character is "root+ vowel "and the root is one of " kha, ga, cha, ja, nya, tha, da, na, tsha, dza".

3. First character is one of 5 prefixes, the number of layer of second character more then or equal to 3.

（1）prefix is "b":

（ i ）second character is "root + superscript or subscript + vowel" and the root is one of "ja, dza, ka, ga, nga, ta, da, sa, za"; notice: when prefix "b" combine with root "da" without subscript.

（ ii ）second character is "root +superscript +subscript + vowel" and root is one of "ka, ga".

（2）prefix is "d",second character is "root +subscript + vowel", the root is one of "ka, ga, pa, ba, ma".

（3）prefix is "m",second character is "root +subscript + vowel" and the root is "kha"or "ga". Special when prefix is "m" and root is "da" without subscript "l".

（4）prefix is “v”, second character is “root + subscript + vowel ”,the root is one of “kha, ga, pha, ba, da” and without subscript “l”.

4. The number of layer of second character more then or equal to 2, subscript is not “w”, second character is one of 9 suffixes except suffix “v”, and only root “ca, cha, zha” combined with long vowel; root can take subscript “w”when root is “ha”and suffix is “ng”; the root can take suffix “s” or “r” when root is “ga”, subscript is “r” and farther subscript is “w”.

C. Normative triple-character syllables

Normative triple-character syllables can be grouped into three types:

1. All the three characters in a syllable are letter

（1）The first letter of this kind of syllable can be any one of those 20 exclusive root letters, second letter ,namely suffix, is one of “*g, ng,, b, m*” and farther suffix is “*s*”

（2）The first letter is one of the five prefix letters, the third letter is one of ten suffix letters:

（ⅰ）if prefix is “*b*”, then root is one of “*ka, ga, ca, ta, da, tsa, zha, za, sha, sa*”;

（ⅱ）if prefix is “*v*”, then root is one of “*kha, ga, cha, ja, tha, da, pha, ba, tsha, dza*”;

（ⅲ）if prefix is “*g*”, then root is one of “*ca, nya, ta, da, na, tsa, zha, za, ya, sha, sa*”;

（ⅳ）if prefix is “*d*”, then root is one of “*ka, ga, nga, pa, ba,ma*”;

（ⅴ）if prefix is “*m*”, then root is one of “*kha, ga, nga, cha, ja, nya, tha, da, na, tsha*”.

2. The layer more then 1 of first character, and its subscript can not be “*w*”, the second character is one of “*g, ng, b, m*”, the third letter is “*s*”.

Special case: when the first character is “*dwa*”, suffix is “*g*” or “*ng*” farther suffix is “*s*”.

3. The first character is one of the five prefixes, the second character has more than one layer, the third character can be any of 9 suffixes:

（1）when the prefix is “b”

（ⅰ） the second character is one of the “ka, ga, ca, ta, da, tsa, zha, za,,sha, sa” combined with a vowel;

（ⅱ）the second character is superscript “*r*” combined with one of the “*ka, ga, nga, ja, nya, ta, da, na, tsa, dza*”; or superscript “l” combined with one of roots “ta, da” ; or superscript “s” combined with one of the roots “ka, ga, nga, nya, ta, da, na, tsa”.

（ⅲ）the second character is root “*ka*” or “*ga*” combined with subscript “*y*” ; root “*ka*” or “*ga*” or “*sa*” combined with subscript “*r*” ; root “*ra*” or “*sa*” combined with subscript “*l*”.

（ⅳ）the second character is superscript “*r*” combined with one of roots “*ka, ga*” and subscript “*y*”; superscript “*s*” combined with root “*ka*” or “*ga*” and subscript “*y*”; superscript “*s*” combined with root “*ka*” or “*ga*” and subscript “*r*”.

（2）prefix is “*v*”:

（ⅰ）the second character is one of the roots “kha, ga, cha, ja, tha, da, pha, ba, tsha, dza” combined with vowel.

（ⅱ）the second character is one of the roots “*kha, ga, pha, ba*” combined with subscript “*y*”; one of the roots “*kha, ga, da, pha, ba*” combined with subscript “*r*”

（3）prefix is “g”:

The second character is one of the roots “ca, nya, ta, da, na, tsa, zha, za, ya, sha, sa” combined with vowel.

（4）prefix is “d”:

（ⅰ）the second character is one of the roots “*ka, ga, nga, pa, ba, ma*” combined with vowel;

（ⅱ）the second character is one of the roots “*ka, ga, pa*” combined with subscript “*y*”, or one of the roots “*ka, ga, pa, ba*” combined with subscript “*r*”.

（5）prefix is “m”:

（ⅰ）the second character is one of the roots “kha, ga, nga, cha, ja, nya, tha, da, na, tsha” combined with vowel;

（ⅱ）the second character is root “*kha*” or “*ga*” combined with subscript “*y*” or “*r*”.

D. Normative quadruple-character syllables

The first letter is one of the five prefixes, and the suffix is one of the letters ‘g’, ’ng’, ’b’, ’m’, and the farther suffix is ‘s’:

(1)Prefix is ‘g’ and the second character is any one of these roots “ca, nya, ta, da, na, tsa, zha, za, ya, sha, sa”, each one combination with vowel as well as.

(2) Prefix is “*d*”:

（ⅰ) the second character is any one of these roots “*ka, ga, nga, pa, ba, ma*”, each one combination with vowel as well as;

（ⅱ）the second character is any one of these roots “*ka, ga, pa, ba, ma*” and subscript is “*y*”, or root is “*ka, ga, pa, ba*” and subscript is “*r*”.

(3) Prefix is “*b*”:

（ⅰ）the second character is any one of the roots “*ka, ga, ca, ta, da, tsa, zha, za, sha, sa*”;

（ⅱ）the second character is a combination of one of the roots “*ka, ga, ca, ta, da, tsa, zha, sha, sa*” with wovel;

（ⅲ）the second character is a combination of superscript“ *r*” with one of the roots “*ka, ga, nga, ja, nya, ta, da, na, tsa, dza*”; or a combination of superscript “*l*” with one of the roots “*ta, da*” ; or a combination of superscript “s” with one of the roots “*ka, ga, nga, nya, ta, da, na, tsa*”

（ⅳ）the second character is a combination of one of the roots “*ka, ga*” with subscript “*y*”; or combination of one of the roots “*ka, ga, sa*” with subscript “*r*”; or a combination of one of the roots “*ka, za, ra, sa*” with subscript “*l*”;

（ⅴ）the second character is a combination of superscript “*r*” or “*s*” with one of the roots“*ka, ga*” and

subscript "*y*"; or a combination of superscript "*s*" with one of the roots "*ka, ga*" and subscript "*r*".

（4）Prefix is "*m*":

（ｉ）the second character is one of the roots "*kha, ga, nga, cha, ja, nya, tha, da, na, tsha*" or a combination of one the of these roots with vowel;

（ii）the second character is a combination of one of the roots "*kha, ga*" with subscript "*y*"; or one of the roots "*kha, ga*" with subscript "*r*".

（5）Prefix is "*v*":

（ｉ）the second character is one of the roots "kha, ga, cha, ja, tha, da, pha, ba, tsha, dza", or is a combination of one of these roots with vowel;

（ii）the second character is a combination of one of the roots "*kha, ga, pha, ba*" with subscript "*y*"; or one of the roots "*kha, ga, da, pha, ba*" with subscript "*r*".

Special explanation about the farther suffix: when the Tibetan language was reformed third time in its long history, it was decided that the farther suffix letter "*da*" should be omitted in writing but still existent in pronunciation, see [3]. Therefore, we will use this farther suffix letter only when we are explaining the grammatical theory and some of the Dunhuang documents and writings that appear on historical monuments set up during Tibetan kingdom. There are 135 triple-character syllables and 186 quadruple-character syllables, such as "*kund,stsold bskyond. bstund*" and so on.

IV. STATISTICAL ANALYSIS OF THE THEORETICALLY NORMATIVE SYLLABLES AND SYLLABLES IN ACTUAL USE

A. Theoretically consistent syllables and syllables in actual use

The generated syllables by the spelling rules of syllable are called theoretically consistent syllables. And the total amount of this type of syllable is 14892, where, 445 single-character syllables, 4985 double-character syllables, 7212 triple-character syllables and 2250 quadruple- character syllables.

The syllables that are obtained from statistical by corpus are called actual syllables which number are 415 single-character syllables, 2475 double-character syllables, 2423 triple-character syllables, 524 quadruple-character syllables respectively , and all together there are 5837.

In the corpus that is covering 6709466 syllables, the subtotals of normative syllables of different length come as following: 1977571 single-character syllables, 2562822 double-character syllables, 1246483 triple-character syllables, 80144 quadruple-character syllables and all together there are 5881912 normative syllables.

All these statistical results indicate: (1) 87.7% of the syllables that are covered within the corpus are normative and a great number of non-normative syllables exist; (2) Of more than fourteen thousand syllables, 39.2% is found to be syllables in actual use. And the theoretical value of these syllables of different length is respectively as following: single-character is 93.26%, double-character syllable 49.65%, triple-character syllable 33.6% and quadruple- character syllable 23.3%.

B. Frequency distributing of actual syllables

In a corpus that is containing 6709466 syllables, times, frequency and accumulative frequency of syllables of different length are studied respectively. As shown in table 2, the syllable of the most frequently is "*pa*" in single-character syllable, 13.2% of all single-character syllables put together. Only 62 single-character syllables can cover about 90.2% and all the other 352 syllables account for less than 9% of the total amount of syllables which are available in Tibetan language. Out of 2489 double-character syllables, only 305 syllables are responsible for the coverage of 90% of texts and the accumulative frequency of all the other 2184 double-character syllables account for less than 9% of the coverage of the texts in corpus. Triple-character and quadruple-character syllables also indicate non-uniformity in the frequency and accumulative frequency. At the same time, we concluded that only 12% of syllables in actual use appear in more than 90% of the texts.

Table 2. Frequency distributing of syllable

Order	Syllable	Times	Frequency%	Accumulative Frequency %
1	པ	258689	13.1814375	13.1814375
62	ཟླ	5181	0.2639968	90.1533127
1	དང	171606	6.5056510	6.5056510
305	ཏོར	1281	0.0485632	90.0341568
1	གཞིས	39529	3.1712427	3.1712427
251	བཟའ	753	0.0604100	90.0003433
1	དཔེགས	7991	9.9708023	9.9708023
61	འབྲངས	203	0.2532941	90.0416718

On the other hand, the distributing of these syllables in different length seems to be non-uniformity in corpus. As shown in table 3, the most frequent single-character syllable is 4.4% of all normative syllables, and the first ten that are ranked most frequent appear 1048985 times and takes up 17.8% of all normative syllables that are covered in the corpus. The most frequent double-character syllable accounts for 2.9% of all normative syllables in statistical results, the first ten most frequent double-character syllables are 10.7% of all normative syllables with the times of occurrence amounting to 629655. The most frequent one among the triple-character

syllables accounts for 0.67% of all normative syllables, the first ten is 4.1% of all normative syllables, occurring 242223 times. The most frequent quadruple-character syllable is 0.14% of all normative syllables, the first ten quadruple-character syllables account for 0.66% of all normative syllables and these syllables occur 38826 times all together.

Table 3. The distributing of first 10 for single-character syllables, double-character syllables, triple-character syllables and quadruple-character syllables

Order	Syllable	Time	Accumulative Frequency %	order	syllable	time	Accumulative Frequency %
1	ཡ	258689	13.1814375	1	གཏིག	39529	5.4904079
2	ལ	145009	20.5703373	2	སོགས	28908	2.3191652
3	བ	122530	26.8138237	3	གས་ཀྱི	28124	7.7466760
4	ད་	91008	31.4511147	4	གཏན	22762	9.5727739
5	ཀྱི	85156	35.7902184	5	རྣམས	21825	11.3237000
6	ས	82021	39.9695778	6	བཞིན	21745	13.0682087
7	ར་	75293	43.8061142	7	རིགས	20950	14.7489376
8	ཤི	65527	47.1450272	8	སེམས	20133	16.3641224
9	ན	62126	50.3106422	9	གཅིག	19312	17.9134407
10	ཀྱི	61626	53.4507790	10	དང་	18935	19.4325142
1	དུ	171606	6.5056510	1	དཔེགས	7991	9.9755325
2	པར	81946	9.6122561	2	ཚུད་	4412	15.4832344
3	ཡིན	70860	12.2985868	3	ཚོམས	4374	20.9435005
4	ནས	57984	14.4967833	4	འཁགས	4084	26.0417442
5	ལས	52169	16.4745312	5	དགོངས	3641	30.5869713
6	མེད	45473	18.1984310	6	བཤགས	3513	34.9724121
7	ཡང	37676	19.6267433	7	དཔང་	2957	38.6637726
8	ཟེད	37375	21.0436440	8	དབྱངས	2823	42.1878548
9	ཡོད	37366	22.4602032	9	དཀྱིལ	2681	45.5346718
10	ཞེས	37200	23.8704700	10	བཞུགས	2350	48.4682846

Besides, the statistical results show that syllables in actual use will change in its level of occurrence across the various corpus, but the high frequency syllables appear to be more stable; The low frequency syllables will have certain change on size of database; The syllables of medium frequency have a slight wave. Only single-character syllable in number changes much little, other syllables have some increase in quantity. This implies the shortage of corpus, such as, the aspect of coverage not enough, and the number be short of corpus.

V Conclusion

In the Tibetan information processing, it is a fundamental duty to do a detail study on normative Tibetan syllables and their in actual use. From the statistical result we can see a certain amount of substandard syllables existing, showing that the universality of mistakes in practice. To summarize, with the help of computer, we can produce theoretically consistent syllables in accordance to the rules involved in the formation of a syllable, obtain distribution of all syllables using statistics, accordingly see the difference between the theoretical value of normative syllables and statistical results of syllables in actual use—syllables in actual use are only 39.2% of the theoretically consistent syllables. For any kind of language information processing, no matter what this language might be, the study of character and word is both the starting point and end-results, therefore, it is an urgent task to build a database containing exclusively normative syllables. As an applied technique, the designed of normative syllables is adopted in Tibetan input system, and the outstanding benefit of this syllable table is that the type of syllable have a function of "orthography" [9].

Acknowledgment

This work was supported in part by The State Language Commission (No: MZ115-69), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Personnel of People's Republic of China, and the tackle key problems in science and Technology Foundation of Gansu province under Grant (No:2GS064-A52-035-04)

References

[1]Weilan Wang. The Frequency rank of Language Unit in Modern Tibetan. Science Technology and Engineering.Vol.4 No.5.2004

[2] Weilan Wang, Wanjun Cheng. Frequency and Information Entropy of Tibetan Character and Syllable. Terminology standardization and Information Ttechnology . 2004.2.

[3] Seduo luosang tsechum gyumsto. Tibetan grammatical theories by Seduo. First edition in Feb.1957, Nationalities Publishing house

[4] Tsedan Xiarong .Detailed Explanation about Tibetan grammar. First Edition in May.1954 Qinghai Nationalities publishing house

[5] Zhade Rengqing Dongzhu. Brief explanation on Tibetan grammatical theories by Zhade renqing dongzhu. Published in 1980 by Qinghai nationalities publishing house

[6] Gama Siduo. Detailed Explanation about Tibetan Grammatical Theories by Siduo. Published in 1982 by Qinghai Nationalities Publishing house

[7] Ruanrao weise. Detailed Explanation about 30 Ode. Published in 1986 by Sichuan nationalities publishing house.

[8] Caidan Xiarong. Tibetan grammatical theories. Published in 1980 by Gansu nationalities publishing house.

[9]Weilan Wang. Intelligent input software of Tibetan. Computer Standards & Interfaces, 2007, 29(4): 462-466.