Calculation analysis on Consonant and character for Corpus study of Gesar epic "HorLing"

DuoLa[1],TashiGyal[2*]

1.Northwest University for Nationalities,Lanzhou,Gansu,China,730030
2.Tibet University,Lhasa,Tibet,China,850000
Duola67@126.com zzxx.77@163.com

ABSTRACT. We made an econometric analysis on consonants and characters after the establishment of Gesar epic classic version HorLing corpus. Firstly, we set up a 2-million-consonants corpus for further verification and comparison of the character frequency of HorLing. Secondly, we established the theory of Tibetan consonants combination rules and Tibetan theory consonants and wish the Tibetan theory consonants coverage on HorLing. By the analysis, we not only understood the status of the Gesar epic consonants and characters, but also clarified the application of Tibetan consonants and characters in the real life.

Keywords: Gesar; Tibetan corpus; consonant and characters

1 Introduction

It is well known that HorLing is the classic version of Gesar epic and the body of the epic. Gesar is the world's longest epic and "living epic". It is still widely talked and sung so far. In spoken and written Tibetan language and dialect differences, it's between spoken and written language. It broke through each dialect barriers and became a universal language. It is important to study the language, which have a restricted view for a nation's way of thinking and attitude of observing.

This paper makes the following contributions: 1) make a qualitative calculation research on vowel sounds; 2) calculate Tibetan theory characters by rule description and software generation; 3) compare with the consonant and characters of Gesar epic, which helps master the language system of the epic; 4) provide a reference data for Tibetan language research and Tibetan information processing. Therefore, it is a very meaningful experiment.

2 Vowels and frequency in HorLing

In ISO/IEC10646 "Tibetan coding Basic Collection" (hereinafter referred to as the "Basic Collection"), the Tibetan is coded according to the dynamic superposition

* Corresponding author

characteristic of Tibetan. It involves Tibetan consonant and component. There are 192 Code bits in Basic Collection, occupying 0F00-0FBF area, 168 coding consonants and components, 24 vacant codes. In Unicode3.0 released in September 1999, 34 Tibetan components and symbols were supplemented. No Tibetan consonant was supplemented in Unicode4.0, but Tibetan coding space extends to OFFF and contains total 256 code bits [1].

According to the consonant, component and symbol in basic collection, we surveyed their distribution and frequency in Gesar epic. In Gesar "HorLing", the consonant, component and symbol occurs 1387431 times and 68 types. There are 3 kinds of symbol appearing in Horling, e.g., "༄", "༅", "༆". The first two are one, but separated in the code. Each of them appears 21 times. The third one is initiative symbol and appears once. There are three punctuations, i.e., syllables, full stop and comma. Among them, syllables occurs the most, i.e., 329814 times and 23.7716% of the HorLing total consonants. Full stop follows, appearing 30145 times and 2.1727%. The third is comma, a total of 12602 times and accounted for 0.9083%.

In addition to the symbols and punctuation, we also investigated four Tibetan vowels contained in "HorLing". The frequencies of four vowels are all shown in Tibetan except vowel "a", which is implied and from which we can understand the cause and focus of Tibetan voice. In the four vowels, vowel "o" appears 63368 times , i.e., the highest frequency, accounted for 6.2442% of corpus amount (except the above punctuation, which occurs 1014827 times consonant and component). Vowel "I" follows, appearing 56968 times and accounted for 5.6136% of total corpus. The next is vowel "u", appearing 48105 times and 4.7402% of total corpus. Vowel "e" is the final one, appearing 46441 times and 4.5762% of the total corpus.

In accordance with the order of traditional grammar, the order should be "i u e o". But from the viewpoint of frequency, the order is naturally "o i u e". The last vowel is adjusted to become the first one. In addition to the above four symbols, we've seen double-layer vowel "o" appearing 22 times, accounted for 0.0022%. The appearance of the vowel "o" is originally from a calling word "ཀྱེ་ཧཽ་ཧཽ".It does not belong to the Sanskrit, but just be used Sanskrit transcription in writing, as "ཀྱེ་བསོ་བསོ" in some versions.

In addition, reversed written vowel "I" appears once. It is not a form of ancient Tibetan's reversed written vowels, but the transcription of Sanskrit words by searching. In the sixth chapter of the first part of HorLing, ChaoTong pretended to be King Gesar, and provoked in the front of the tribe Hore camp. There is a sentence in the lyrics"ལྷག་ཚན་འདི་ལ་ཚ་མོ་ཡོད༎ སྔོན་པ་འཛམ་བུ་ཤྀགླ་ཤིང་ཡིན༎"[2]. The reversed written vowels combinations are the name of the tree of life in South ZhanBu continent, which are loanwords and derived from Sanskrit. They are occasionally used in the Tibetan text, but rarely in spoken Tibetan. It also suggests that Gesar epic language exists between spoken language and text, or Gesar epic language is an organic combination of text and spoken language. At the same time, the reversed written vowel is not ancient Tibetan, which means Gesar epic language is the common language and does not contain a lasted synchronic language. It may be one of the reasons for epic "Gesar" as the only "living epic".

3 The consonant and component and their frequency in Horling

The top ten are "ས་ག་ད་ར་མ་བ་ང་ན་ལ་འ" except individual punctuation and vowel in Horling. "ས"appears 84921 times, the highest frequency, accounted for 8.3680% of the corpus amount (the corpus contains 1014827 times of consonant and component except syllables and punctuations); "ག"appears 72025 and accounts for 7.0973% of the total amount of corpus. To be sure, the frequency of "ག" does not include "ག"as a root when combined by ko can and dok can. Frequency of root of "ག" statistic separately (the following consonant acing as other similar function are the same). "ད" appears 69418 times, accounted for 6.8404% of the total. "ར" 63281 times, 6.2356%; "མ" 48392 times, 4.7685%; "བ" 48320 times, 4.7614%; "ང" 46583 times, 4.5902%; "ན" 45724 times, 4.5056%; "ལ" 40112 times, 3.9526%; At the end of the "འ", a total of 37850 times, close to 40000, accounted for 3.7297%.

Fortunately, if we adjust these 10 characters according to the order of traditional grammar, the result is very surprising. The traditional order of the ten words is "ག་ང་ད་ན་བ་མ་འ་ར་ལ་ས", which is just the ten latter root of Tibetan character. Therefore, how could the latter root be high frequency consonant in consonant and component, and what is its primary cause? We will deeply discuss that later combined with character status.

Syllable, full stop and four vowels, 10 latter root, a total of 16 high-frequency consonants and punctuation, and then is the superimposed consonants and other letters. The highest frequency is the first two of the four dok can in Tibetan, i.e. ya and ra, The frequency of "ྱ"and "ྲ", "ྱ" are of 29717 times, 2.9283%; frequency of "ྲ" appear 20773 times, accounted for 2.0469%, the frequency of which is below 20000 and more than 30000. The frequency of the third dok can is"ླ", accounted for 0.8404%, significantly lower than the first two, less than 1%. Even so, "ླ" is not only dok can, and sometimes acts as root of other ko can, such as "ཟླ". The fourth dok can" ྭ"only appear 649 times, accounted for 0.0640%, which ranked 51 in frequency of 62 consonants and components.

In addition, "ར་ལ་ས" has three ko can in Tibetan. Unfortunately, due to not taking traditional way for encoding, but making dok can as a key coding, statistics on frequency of the three ko can cannot adopt the same method, which had to be observed in the part of discussion of character. However, this encoding way provided us an opportunity to research ko can combination roots. There are 16 ko can roots. 3 roots are more than 8000 high frequency, respectively is "ྒ", "ླ", "ྟ". Among them, frequency of " ྒ" is 10831 times, 1.0673%; frequency of "ྟ" is 8020 times, accounting for 0.7903%. "ླ" has talked about as ko can, here is mainly aimed at its root function, these three characters could act as common root for three ko can.

In conclusion, there are 1387431 times of consonants, components and punctuation in "HorLing" of "Gesar" epic. There are 62 types of consonants and components except six punctuation, covering 1014827 consonants and components. We could see that only 6 consonants appear in 15 vowel consonants of Horling according to Basic collection. In addition to once Sanskrit reversed written vowel, the rest can be considered to be Tibetan vowels, coverage of 100%. Appeared in the top 10 are 10 latter

roots in the scope of letters. They represent not only latter roots, but appear as roots and act as latter root. The other five could also act as previous root. "རྣལ་ས" could be used as ko can, but they cannot used as ko can roots. The diversity of the role makes them stand out in more than 1 million consonants and become a top ten high frequency consonants. More impressive is that the cumulative frequency of these 10 consonants and 4 vowels achieved 76.0236%. That is to say, 1014827 consonants in the full text of Horling except syllables and punctuation, 76% of the 14 consonants were repeated and composed in the various roles.

The highest level of the art is not to make the simple things complicated, but to describe the complicated things by the most simple and fluent language, as is not only the art and science. If this view is correct, we could say that the epic of "Gesar" is a perfect combination of art and science. We could also make this question, of course, the characteristics of the phonetic writing is to realize complex expression by a few letter combinations. There is no doubt that it should be acknowledged. What make us surprising is that 30 letters of alphabet writing should only use 10 of them and reached 76% of expression. This work is a miracle.

In 168 coded consonants version of "Basic collection", there has 77 consonant characters [3], including Tibetan single consonant, 49 dok can root character, 28 Sanskrit consonant character and dok can characters. In the epic Gesar, statistics of consonant and components of "HorLing". There are 49 pureTibetan characters except seven vowels and six Sanskrit characters. It shows that in the level of consonant and component, "Gesar" epic" HorLing" covers 100% of the modern Tibetan consonant and component, effectively prove "HorLing" has important value as a template for Tibetan language research.

4 About the character

Tibetan has 206 basic character in traditional grammar (ལྷག་ཆྱི་ཡི་ཀ)[4] if "body" understood as character, the statistics might not consider other factors. But the basic character is put forward, which gives us a consideration and evolves “character” .For the attention of the "character" and put forward again of this concept due to the rise of Tibetan information technology, the problem of character were raised at the beginning of encoding for Tibetan.

In terms of Chinese characters, a word can be classified as a character, such as 6763 Chinese characters were collected in GB 2312, including 3755 primary/first-level Chinese characters, 3008 secondary characters, which basicly meet the requirements of regular users of computer processing of Chinese characters, and its coverage to mainland China 99.75% of Chinese characters operating frequency. From which we can see that every word is a syllable unit, at the same time it is a character.

5 The theory about Tibetan character

For the Tibetan language, there is a big difference between characters, a syllable is not necessarily a character, a word at the same time sometimes could be a syllable. As mentioned above, a Tibetan syllable is consists of seven parts at most . Although it also like Chinese is just one syllable, but not take it as a character. In fact, Tibetan cannot include previous root, latter root and plus latter root as character , just calculate single character and vowel, and vertical superposition (ko can,dok can).

We need to know how many character in Tibetan, how many characters in voluminous ancient books and the epic? Three categories and 33 ko can in Tibetan, four categories and 40, four vowels, 30 roots, these combined according to the rules of word groups referred to as lus kyi yi e,namely,body word or character.

In order to study Tibetan character, according to the theory of Tibetan voice, we made a rule description and generated in program. In this way,

there are 454 Tibetan characters in the generating program. Including the translation of words: རྣ་རྩ་ཞུ, the three actually belong to the category of Sanskrit long vowels. In addition, five new words namely ཐ་ཐི་ཐུ་ཐེ་ཐོ and 39 DunHuang[5][6] literature and stone ancient Tibetan: ཀྱ་ཀྲ་འཡ་ཁྲ་ཁྱ་ཁྲི་ཁྱི་ཁྲོ་ཁྲུ་ཁྱུ་ཁྲེ་ཁྱེ་ཁྲོ་ཁྱོ་རྒྱ་རྒྲ་རྒྱི་རྒྲི་རྒྱུ་རྒྲུ་རྒྱེ་རྒྲེ་རྒྱོ་རྒྲོ་རྒྱུ་རྒྲུ་རྒྱེ་རྒྲེ་རྒྱོ་རྒྲོ་རྣ་རྩ་ཞུ Among them, the "ཇ" character pronunciation should be "ja",it is a person's name, suspected that the strokes weren't connect well when sketching,

"ཀྵ" written like "ཤྐ", actually should be "ཀྵ". In this way, the total number of Tibetan character should be 498. Of course, there are also some superposition of latter root and plus latter root in the DunHuang literature and ancient Tibetan stone tablets, this part actually are superposed according to the need of writing space at that time. Ultimately, they are still latter root and plus latter root, it will not have any change in voice except Non-standard writing .The 498 Tibetan character, like 400 no-tone basic Chinese syllable, which could be regarded as the foundation of the Tibetan voice systems or basic voice.

But we need to know, if only count the modern Tibetan characters, then minus three Sanskrit reversed character "རྣ་རྩ་ཞུ" in the 454 characters ,and add five new character "ཐ་ཐི་ཐུ་ཐེ་ཐོ"[7]. In this way, the number of simple modern Tibetan character should be 456.

6 Character and frequency of HorLing

As the world's longest epic "Gesar" ,how many basic voice used when talking and singing, or how many characters for recording in Horling Part one and two? If we could make this clear, then we could understand the reality of the Tibetan voice system probably.Based on this understanding, we counted the number of characters of "HorLing"part one and two, there are altogether 445 characters.

Among them , there are 34 Sanskrit, it respectively is ཀ་ཁ་ག་ང་ཅ་ཆ་ཇ་ཉ་ཏ་ཐ་ད་ན་པ་ཕ་བ་མ་ཙ་ཚ་ཛ་ཝ་ཞ་ཟ་འ་ཡ་ར་ལ་ཤ་ས་ཧ་ཨ་ཀྵ་ཇྷ་ཉྩ་ཏྲ་ཐྲ་ཤྲ་ཧྲ་ཀྲ་ཁྲ་གྲ་ཕྲ་བྲ་མྲ་ལྲ་ཀྱ་ཁྱ་གྱ་ཕྱ་བྱ་མྱ་ལྱ་ཀྭ་ཁྭ་གྭ་ཕྭ་བྭ་མྭ་ཀྷ་ཁྷ་གྷ་ཕྷ་བྷ་མྷ་ཀྵ་ཇྷ་ཉྩ་ཏྲ་ཐྲ་ཤྲ་ཧྲ་ཀྲ་ཁྲ་གྲ་ཕྲ་བྲ་མྲ་ལྲ་ཀྱ་ཁྱ་གྱ་ཕྱ་བྱ་མྱ་ལྱ་ཀྭ་ཁྭ་གྭ་ཕྭ་བྭ་མྭ་ཀྷ་ཁྷ་གྷ་ཕྷ་བྷ་མྷ་ཨ་ཨི་ཨུ་ཨེ་ཨོ་ཨཾ་ཨཿ་ཨྃ་ཨྂ་ཨ྄ ,so pure Tibetan left 411.There are no five new words in the epic, three translation character and the reversed-written vowel ancient Tibetan char-

acter, and appear some Sanskrit categories, but the frequency of the Sanskrit is very low, just appear while reciting om mani padme hung and ali's four great rivers, and several animal and plant names such as "water lotus", "lion"etc. 34 Sanskrit and frequency table (table 1) as follows:

Table 1. Frequency table of the Sanskrit and frequency

	charac-	time	frequen-	No	charac-	time	frequen-	Amount
1	ཨ	142	0.0205	18	ཀླ	2	0.0003	69335
2	ཨཱ	30	0.0043	19	ཁ	1	0.0001	69335
3	ཧཱ	22	0.0032	20	ཀཱཾ	1	0.0001	69335
4	ཤྲཱ	17	0.0025	21	ཀྵ	1	0.0001	69335
5	ཨཱཾ	13	0.0019	22	ཀྟ	1	0.0001	69335
6	ཀཱ	12	0.0017	23	ཀྐ	1	0.0001	69335
7	ཀ	7	0.0010	24	ཀྱཱ	1	0.0001	69335
8	ཀྱ	6	0.0009	25	ཀྲཱ	1	0.0001	69335
9	ཀཾ	5	0.0007	26	ཀྲ	1	0.0001	69335
10	ཨཱཿ	5	0.0007	27	ཀཱ	1	0.0001	69335
11	ཨཱྀ	5	0.0007	28	ཀྀ	1	0.0001	69335
12	ཨཿ	4	0.0006	29	ཀྭཱཾ	1	0.0001	69335
13	ཀྵི	3	0.0004	30	ཀྵྣ	1	0.0001	69335
14	ཨྃ	2	0.0003	31	ཀྱཱ	1	0.0001	69335
15	ཀྃ	2	0.0003	32	ཀྵཱ	1	0.0001	69335
16	ཀྀ	2	0.0003	33	ཨཱྀ	1	0.0001	69335
17	ཀཱ	2	0.0003	34	ཨཱ	1	0.0001	69335

(Note: not including in the total number of syllable point)

As you can see, in 34 Sanskrit character, a total of 17 character appear twice or more, the rest is only appear one time. The highest frequency in Sanskrit is "ཨ", a total of 142 times, in fact the character and its combination are both "ཨ་དུང", it is a character with different form of "དཔའ་རྩལ (hero)".

We already know that the number of modern Tibetan character is 456, which appear 411 times in "HorLing", there are still 45 Tibetan character does not appear, and sill 40 characters did not appear in the epic except 5 new character "ཐྱ་ཐྱི་ཐྱཱ་ཐྱཱི་ཐྱོ". 411 Horling character accounted for 90.1316% in modern Tibetan characters, there are nearly 10% of the them has not been covered. If some high-frequency character did not appear, then typicality of the HorLing language will be questioned. Therefore, we need to know which character does not appear, and the function of those does not appear.Through comparison and selection, we found that the 40 characters are: "ཀྱ་ཡེ་ཤྲེ་ཤྲ་

ཐྲྀ་ཐྲྀ་ཏེ་ཐྲི་ཐྲ་ཐྲུ་ངི་འེ་ཇ་ཇུ་ཀུ་ཀྱ་ཀྲ་ཀྲི་ཀྲེ་ཀྲོ་ཀྲྀ་ཀྲཱ་ཀྲཱི་ཀྲུ་ཀྲཱུ་ཀྲྀ་ཀྲཱྀ་ཀླ་ཀླི་ཀླུ་ཀླེ་ཀློ་ཀླྀ་ཀླཱྀ་ཀྵ་ཀྵི་ཀྵུ་ཀྵེ་ཀྵོ་ཀྵྀ་ཀྵཱྀ་ཀྵླ་ཀྵླི་ཀྵླུ་ཀྵླེ་ཀྵློ་ཀྵླྀ་ཀྵླཱྀ་ཀྵྱ་ཀྵྲ་ཀྵྲི་ཀྵྲུ་ཀྵྲེ་ཀྵྲོ་ཀྵྲྀ་ཀྵྲཱྀ་ཀྵྭ་ཀྵྭི་ཀྵྭུ་ཀྵྭེ་ཀྵྭོ་ཀྵྭྀ་ཀྵྭཱྀ་ཀྵྵ་ཀྵྵི་ཀྵྵུ་ཀྵྵེ་ཀྵྵོ་ཀྵྵྀ་ཀྵྵཱྀ་ཀྵྺ་ཀྵྺི་ཀྵྺུ་ཀྵྺེ་ཀྵྺོ་ཀྵྺྀ་ཀྵྺཱྀ་ཀྵྻ་ཀྵྻི་ཀྵྻུ་ཀྵྻེ་ཀྵྻོ་ཀྵྻྀ་ཀྵྻཱྀ་ཀྵྼ་ཀྵྼི་ཀྵྼུ་ཀྵྼེ་ཀྵྼོ་ཀྵྼྀ་ཀྵྼཱྀ་ཀྵ྽་ཀྵ྽ི་ཀྵ྽ུ་ཀྵ྽ེ་ཀྵ྽ོ་ཀྵ྽ྀ་ཀྵ྽ཱྀ་ཀྵཾ་ཀྵཿ་ཀྵྃ་ཀྵ྄", in general, the 40 characters are not commonly used, and even some could be called "death character", never used, such as "ཀྵ"[8].

The interpretation to meaning item of each character and function of character building,we query to the 30 text file of "reference corpus", total size is 4.11 MB, 2.15 million characters. In which 32 words never appear; the highest frequency appear 99 characters, belongs to the ancient Tibetan "ཞི (people)", following is 6 times, 3 times, one time.

It shows that the 40 character is not commonly used, most of them never appear in the 2 million corpus. The ancient Tibetan accounted for the proportion , in addition to the individual character, most is uncommon character. Therefore, it is reasonable that they did not appear in HorLing , because the characteristics of HorLing language is very popular, a universal language.

About the new character "ཐྭཱ་ཐྭི་ཐྭུ་ཐྭེ་ཐྭོ" does not appear in "HorLing". For the five new character, there are two statements. One is the five characters are new combination according to the need of translation after liberation, to translate some"f"consonant such as "Leifeng", "plane" and "Mei lanfang" should adpot five characters; Another is that these five characters has been used in the period of the fifth dalai lama (the early qing dynasty). Anyway, the characters might be created lately from the situation of non-appear in Holing corpus.There is no this kind of pronunciation in the folk oral Tibetan. Moreover, so far, there hasn’t local characters formed by these characters except translation need, related basic voice replaced with pa and pha, no"f"consonants or initials words.

We already know that 411 characters in"HorLing"accounted for 90.1316% of modern Tibetan characters, also include some Sanskrit common used characters, to a certain extent, it reflect the reality of the usage of Tibetan. It is inevitable that some Sanskrit exists in Tibetan regularly ,therefore, it is necessary to taken into account of the part of the reversed written Sanskrit while coding for Tibetan . In the real life level, it also reflects the Tibetan is a universal buddhist nation and it has the very deep influence as the birthplace of Buddhism.

The top ten high frequency character are "ས་གི་ད་ང་ན་མ་ར་བ་ལ་འ" in the epic, we are very familiar with these ten characters. The top 10 in statistical consonants are "ས་གི་ད་ར་མ་བ་ང་ན་ལ་འ", it is different in order compared with characters. Actually, they are just ten latter roots if we reorder these ten characters,that is, "གི་ང་ད་ན་བ་མ་འ་ར་ལ་ས".

We know that the total frequency the character is 693356 times, the cumulative frequency of top 10 character is 391069 times, 56.4023% of the total number of characters. Detailed in the following table (table 2) :

Table 2. frequency of top 10 characters list

No.	character	times	frequency（%）	cumulative times	cumulative frequency（%）	Amount
1	ས	53669	7.7405	53669	7.7405	693356

2	གྱ	48251	6.9591	101920	14.6995	693356
3	ད་	46848	6.7567	148768	21.4562	693356
4	ད	44149	6.3674	192917	27.8237	693356
5	ན་	44031	6.3504	236948	34.1741	693356
6	མ	34726	5.0084	271674	39.1825	693356
7	ར	34056	4.9118	305730	44.0942	693356
8	བ	31894	4.5999	337624	48.6942	693356
9	ལ	27888	4.0222	365512	52.7164	693356
10	ཞ	25557	3.6860	391069	56.4023	693356

In characters, it is still a"ས" in the first place, the second and third places are "གྱ" and "ད་" same as characters statistics layout, and then change on the ordinal, but on the whole, still the 10characters.

In the top 100chracters, mainly is function word, such as five the auxiliary case are in the top 50. They are, "ཏེ" ranks 12th, appear 10408 times, frequency is 1.5011%; "ཡི" in the 18th, 5054 times, the frequency is 0.7289%; "ཀྱི" in the 33th, 3171 times, frequency is 0.4573%; "ཅེ" in the 44th, 2568 times, frequency is 0.3704%; "ཀྱི" is 46th, 2417 times, frequency is 0.3486%. The frequency of ordering shows that it is obvious for the adhesion phenomenon(function word cohere with national word) of the case.

The cumulative frequency of the first 100 characters is up to 625379 words, accounted for 89.94% of total corpus, the lowest frequency is 945 times. The 100 characters appear 1033 times except the end of the last three characters. So they can be called thousand-times character.

From the number of 101to 200, character root has taken main status. The highest frequency of this phase is 935 times, the lowest is 265 times of the 200th character. Which is relatively balanced, unlike the frequency of the first 100 characters, the first appeared 53669 times, and 100th appear 945 times, they are differ more than 52000 times.

The cumulative frequency of the first 200 characters is up to 675826 times, accounted for 97.47% of the total number. From the number of 101 to 200, mostly is superposition character, followed by consonant and vowel, illustrate the major component of Tibetan alphabet combination in this frequency. The character without superposition and vowel is only one "ཿ", its number is 107, still in the first place, appear 872 times, frequency is 0.1258%.

The frequency and distribution of 447 characters shows that the reflects the usage and distribution of Tibetan characters in epic, on the other hand it also reflects a general situation of the entire Tibetan character. All 447 characters including Sanskrit, there are 29 of which frequency is only once, 10 of which frequency is 2, seven of which frequency is 3, five of which frequency is 4, three of which frequency is 6, only one of which frequency is 7, two of which frequency is 8, there of which frequency is 9, two of which frequency is 10, there of which frequency is 11, there of which frequency is 12, there of which frequency is 13.As you can see from these frequency, 2-

3characters distributed for each level in low-frequency ones equably. Throughout the 447 characters, there are 115 characters which frequency are below 30.The rest 33 Sanskrit except "ཨ" are mixed In this frequency channel. It can be seen that the character of this frequency channel is not "uncommon character".

7 Conclusion

"HorLing", commonly known as "HorLing wars "or" the war of HorLing", mainly describes the war events between tribe of Hore and Ling, shows the changes of the Tibetan tribal society. Mr Qian mintz wrote in his article: "HorLing wars is a historical picture scroll of Tibetan hero's grand, majestic and grand, it is permeated with enthusiasm upward, positive enterprising and fight passion. This is very similar to the Iliad"[9] He also asserts that "Tibetan HorLing wars is the Chinese nation's "Iliad""[10].

It is earlier to study of A Dream of Red Mansions by corpus method, which inspired us for the first time to study Gesar epic by establishing Tibetan tagging corpus. Though the study we not only get the number of Tibetan theory character, but also verified the frequency of character in the practical application, and classified the characters. Moreover, we verified the 40 low-frequency characters in a larger range (2 million characters) the verification of the corpus to further confirm the reliability of the frequency of characters in "HorLing", it makes sense.

ACKNOWLEDGEMENTS

This research project was financially supported by the National Natural Science Foundation of China (NSFC)(No. 61262053) and The ministry of education philosophy and social sciences research project for major project(No. 13JZD028), the National Natural Science Foundation of China (NSFC)(No. 61163043).

REFERENCES

1. Di Jiang,Cong Jun Long.:Study on Tibetan character, Social sciences academic press,Aug,pp.88-89(2010)
2. Qian mintz, HorLing wars, the Iliad, Foreign literature,pp.125-251 (1986)
3. Di Jiang,Cong Jun Long.: Study on Tibetan character, Social sciences academic press,Aug, pp.77-78(2010)
4. Caidanxiarong.:The Tibetan language grammar,Qinghai minorities press, pp.25-26(1998)
5. Edited by Luo Bingfen, Tubo Medical Literature Essential, Ethnic Publishing House, pp. 127-300(2002)
6. Edited by Chenjian, Wangyao.:Duanghuang Manuscript Tibetan Literature, ethnic publishing house , pp.12-321(1983)
7. Compiled by Tsering Tso etc.: Atomic Physics (Tibetan-Chinese Languages), Gansu Ethnic Publishing House, pp.7-29(2005)

8. Compiled by Zenlha Ngawang Tsutrin, Ancient Tibetan Dictionary, Ethnic Publishing House, pp.15-16(1997)
9. Qian mintz, HorLing wars, the Iliad, Foreign literature,pp.103-105 (1986)
10. Qian mintz, HorLing wars, the Iliad, Foreign literature, pp.114-120(1986)