Zipf's Law and Statistical Data on Modern Tibetan

Huidan Liu
Institute of Software,
Chinese Academy of
Sciences, Beijing,
China, 100190
huidan@iscas.ac.cn

Minghua Nuo
Institute of Software,
Chinese Academy of
Sciences, Beijing,
China, 100190
minghua@iscas.ac.cn

Jian Wu
Institute of Software,
Chinese Academy of
Sciences, Beijing,
China, 100190
wujian@iscas.ac.cn

Abstract

In this paper, a large scale modern Tibetan text corpus is built, which includes about 190 thousands documents, 67.21 million words, 93.66 million syllables in total. Based on the corpus, statistics are made in several language units in different granularities. Statistical data show that : a syllable has 3.26 letters or 2.20 super characters in average, while a sentence has 75.40 letters or 63.14 super characters. The top 10 super characters, syllables, words take up 66.3156%, 16.5556%, 24.6415% of the corpus respectively. Curves for the n-gram frequency-rank list of super chars, syllables and words are plotted. It shows that when all the n-gram phrases for $n = 1, 2, \ldots, 5$ are put together and sorted by frequency in descending order, the frequency-rank curves in log-log axes can be fitted well by a straight line for the unit of syllable and word respectively. But for the unit of super character, we didn't find a curve that can be fitted well enough by a straight line even if we combine all the n-grams for $n = 1, 2, \ldots, 10$.

1 Introduction

The statistical property is the natural property of a language. In recent tens of years, people made statistical analysis on Tibetan characters or syllables. But it's difficult to make statistics on larger language units such as word and n-gram word phrases, especially on a large scale corpus. There are two reasons resulting in the difficulty. First, as Tibetan is a resource poor language, it's hard to build a large scale Tibetan text corpus. Second, Tibetan word segmentation technology is not well developed even until now.

In this paper, we report our word on the statistics on Tibetan based on the language units such as character, syllable, word and their n-gram pairs on a large scale corpus. The remainder of the paper is organized as follow. Tibetan language units are introduced in Section 2. We recall the related work in Section 3. In Section 4, the methods which is used to build the corpus and to segment Tibetan text into language units are described in detail. We make statistics on the corpus and list the most frequency Tibetan language units and test Zipf's law respectively in Section 5. Section 6 concludes this paper.

2 Language Units of Tibetan

Generally speaking, Tibetan is a alphabetic writing system. But there is a unit larger than letter but smaller than syllable, which is different from other language such as English and Chinese. Meanwhile, people have used different terms (in English) to express the same unit or the same term to express different units. So we must make a clarification in this Section.

2.1 Letter, Character and Super Character

There are 30 consonants and 4 vowel signs in modern Tibetan. Several other consonants and vowel signs are also used in Tibetan text to transliterate Sanskrit script. There are only 4 vowel signs (writing) for the vowels (reading) /e/, /i/, /o/ and /u/, but there isn't any signs for the vowel /a/, so every consonant has

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 322–333, Dublin, Ireland, August 23-29 2014.

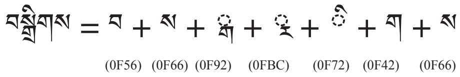

བསྒྲིགས = བ + ས + ྒ + ྼ + ི + ག + ས
(0F56) (0F66) (0F92) (0FBC) (0F72) (0F42) (0F66)

Figure 1: Tibetan encoding schema with small unit used in ISO/IEC 10646.

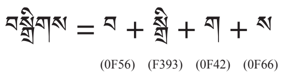


Figure 2: Tibetan encoding schema with large unit used in GB/T 20542.

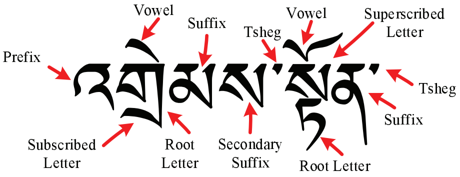
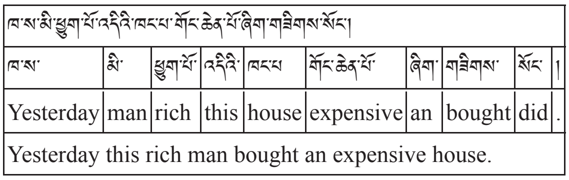


Figure 3: Structure of a Tibetan word.



ཁ་ས་མི་ཕྱུག་པོ་འདིའི་ཁང་པ་གོང་ཆེན་པོ་ཞིག་གཟིགས་སོང་།									
ཁ་ས་	མི་	ཕྱུག་པོ་	འདིའི་	ཁང་པ	གོང་ཆེན་པོ་	ཞིག་	གཟིགས་	སོང་	།
Yesterday	man	rich	this	house	expensive	an	bought	did	.
Yesterday this rich man bought an expensive house.									

Figure 4: A Tibetan sentence.

an inherent vowel /a/. Other vowels can be indicated using a variety of diacritics which appear above or below the main letter. Each of these consonants and vowel signs is called a "letter".

In Tibetan encoding schema used in ISO/IEC 10646 and Unicode standard (Consortium, 2013), each Tibetan consonant has two or even more code points to denote its normal form or subjoined form, or other variant forms which are only used in very special context. Each variant form is called a "character" corresponding to a code point in ISO/IEC 10646. In Figure 1, seven characters form a Tibetan syllable. Note that three consonants and a vowel sign are clustered.

Different from the encoding schema with small unit used in ISO/IEC 10646, in Chinese notional standard GB/T 20542 and GB/T 22238 on Tibetan coded character set and some legacy Tibetan encodings, another encoding schema is used. In this schema, the cluster of consonants and vowel sign in Figure 1 is assigned only one code point. Figure 2 shows the schema. The encoding unit shown in Figure 2 is called "字丁 " (Zi Ding) in Chinese but the Chinese term doesn't have an exact English translation, and we call it "super character" or "super char" briefly in this paper.

2.2 Syllable, Word and Sentence

A syllable contains one or up to seven character(s). Syllables are separated by a marker known as "tsheg", which is simply a superscripted dot. People sometime use the Chinese term "字 " (Zi, exactly a character in Chinese script) to denote "syllable". The term "字 " is often translated to "word" in English. But "word" mainly used to express a larger language unit as an item in the vocabulary. So we use the term "syllable" in this paper, and take "word" as a larger language unit which is made up of one or more syllables and has meanings.

Note that in Tibetan "tsheg" is used as the delimiter between two syllables. But there is no another delimiter to mark the boundary between two words. Thus there is a lack of word boundaries in Tibetan. Figure 3 shows the structure of a Tibetan word which is made up of two syllables and means "show" or "exhibition".

In Tibetan text, some monosyllable words, including "འི", "ས", "ར", "འང", "འམ", "འོ" (We call them abbreviation markers (AM) in this paper), can glue to the previous word without a syllable delimiter "tsheg", which produce many abbreviated syllables. For example, when the genitive case word "འི" follows the word "རྒྱལ་པོ" (king), we don't put a "tsheg" between them and get the fused form "རྒྱལ་པོའི" (king[+genitive]). The existence of abbreviated syllables contributes to the difficulty to segment Tibetan sentence into words.

Tibetan sentence contains one or more phrase(s), which contain one or more words. Another marker known as "shed" indicates the sentence boundary, which looks like a vertical pipe. Figure 4 shows a Tibetan sentence and its translation in English.

3 Related Work

In the early 1930s, G. K. Zipf pointed out a statistical feature of large language corpora (both written texts and speech streams) which, remarkably, is observed in many languages, and for different authors and styles (Zipf, 1935). He noticed that the number of words $w(n)$ which occur exactly n times in a language corpus varies with n as $w(n) \sim 1/n^{\gamma}$, where the exponent is close to 2, which results in the well known Zipf's law. The general form of Zipf's law states that:

$$y = f(r) = \frac{C}{r^{\alpha}} \tag{1}$$

where α is a positive parameter close to 1.

Zipf showed that, by and large, his law held for words, syllables and morphemes. Consequently, it is natural to ask if the law also holds for pairs of words. Egghe devised a mathematical argument that it, in fact, does not, but that the exact relation can be approximated by a power law (Egghe, 1999). He extended his investigations to parts of words, namely to the study of N-grams (Egghe, 2000).

Zipfs law was the source of a lively debate related to the structure of DNA. It was claimed (Mantegna et al., 1994) that Zipf's law shows the difference between coding and non-coding DNA as non-coding (so-called junk) DNA fits Zipf's law much better than coding DNA. This would mean, according to the authors, that non-coding regions of DNA may carry new biological information. Yet, this does not mean that junk DNA is a kind of language. Other scientists (Chatzidimitriou-Dreismann et al., 1996), however, have shown that this distinction is not universal and lacks all biological basis.

Zipf's law has been tested on the Internet. It turned out that popularity of Internet pages is described according to Zipf's law. This fact can be used to design better cache tables (Masaki and Takahashi, 1998; Breslau et al., 1998; Adamic and Huberman, 2002). Zipf's studies on city sizes still lead to new developments in geographical and economical studies (Gabaix, 1999a; Gabaix, 1999b; Okuyama et al., 1999; Ioannides and Overman, 2003; Soo, 2005; Soo, 2007).

Back to text, Li (1992) found that the distribution of word frequencies for randomly generated texts is very similar to Zipf's law observed in natural languages such as English (Li, 1992). Ha et al. (2002) investigated the law for two languages English and Mandarin and for n-gram word phrases as well as for single words. The law for single words is shown to be valid only for high frequency words. However, when single word and n-gram phrases are combined together in one list and put in order of frequency the combined list follows Zipf's law accurately for all words and phrases, down to the lowest frequencies in both languages. The Zipf curves for the two languages are then almost identical (Ha et al., 2002).

In recent years, researchers also made statistics on Tibetan. Jiang and Dong (1994) made statistics on the length and different structural mode of Tibetan syllables, and counted up the number of initial clusters and finals of Tibetan syllables, as well as the number of Tibetan letters at different positions in syllables (Jiang and Dong, 1994; Jiang and Dong, 1995). In a further research Jiang (1998; Jiang and Kong (2006; Jiang and Long (2010) , they made statistics on Tibetan letters, and found that the 1th order and 2nd order entropy of Tibetan is 3.9913 bits and 1.2531 bits resplectively (Jiang, 1998), while on super character they are 4.82 and 3.12 (Jiang and Kong, 2006; Jiang and Long, 2010). Wang and Chen (2004) made similar research to calculate the frequency and information entropy of Tibetan character and syllable based on a corpus of $20,000,000$ characters, and discovered that the most frequent 703 Tibetan syllables cover 90% of the corpus (Wang and Chen, 2004). She also presented the research on the frequency-rank relation of Tibetan super character and syllable, and found that the distributions follow Zipf's law too (Wang, 2004). But no further research is reported on whether she tests Zipf's law on larger language units of Tibetan. Other researchers also made statistics on Tibetan syllable's structural mode based a static corpus such as syllable list or dictionary (Gao and Gong, 2005; Ai et al., 2009). Lu et al. (2003) presented the theories and approaches to calculate the frequencies of Tibetan characters, pieces, syllables and words based on a large scale Tibetan corpus including about $40,000,000$ syllables (Lu et al., 2003). However, a large part of the corpus they used are Buddhist literatures and the work can't be done well without a pragmatic Tibetan word segmentation tool (Chen et al., 2003a; Chen et al., 2003b; Jiang, 2006; Jiang and Kong, 2006; Sun et al., 2009; Sun et al., 2010; Lu and Shi, 2011; Liu et al., 2012a).

At present, people already find methods to build a large scale corpus from Tibetan web sites with low cost. Liu et al. (2012b) presented their method to extract the title, content, author and other useful information of articles from several news and broadcasting web sites (Liu et al., 2012b). It's not a difficult work to implement a pragmatic Tibetan word segmentation tool based on the former researches (Chen et al., 2003a; Chen et al., 2003b; Sun et al., 2009; Sun et al., 2010; Liu et al., 2012a) So it's time to make statistics on the frequency distribution of larger language units such as word and n-gram word phrase for Tibetan to see whether they follow Zipf's law.

4 Methods and Corpus

We present our methods to build the corpus and to segment Tibetan text into different units mentioned above.

4.1 Building a Large Scale Tibetan Text Corpus

Previously Liu et al. (2012b) proposed an approach to build a large scale text corpus for Tibetan natural language processing. We adopt the method to build our corpus. we crawled eight Tibetan websites which mainly focus on news and broadcastings. Topic pages but hub pages are selected with a rule based method by checking the url. We analysed the layout structure mode of each web site and built templates to extract topic title, publishing date, author, topic content and some other topic related informations.

Consequently, a large scale Tibetan text corpus is built, which includes about 190 thousands documents, 67.21 million words, 93.66 million syllables and 265 million super characters in total. The sources and scales in different units are shown in Table 1.

♯	source	♯document	♯sentence	♯word	♯syllable	♯super character
1	http://tb.chinatibetnews.com	74,632	1,419,967	26,648,803	37,633,467	108,010,715
2	http://tb.tibet.cn	13,348	331,022	4,288,187	5,872,524	16,388,242
3	http://ti.gzznews.com	8,084	281,405	3,518,918	4,763,097	13,301,408
4	http://ti.tibet3.com	26,631	725,669	9,186,980	12,634,804	35,595,345
5	http://tibet.people.com.cn	29,797	833,221	9,323,838	12,908,542	35,328,443
6	http://www.qhtb.cn	20,616	575,242	7,908,508	10,913,097	31,200,465
7	http://www.tibetcnr.com	9,559	278,681	3,272,274	4,624,878	13,114,130
8	http://xizang.news.cn	7,707	187,423	3,062,419	4,307,175	12,258,911
Total		190,374	4,632,630	67,209,927	93,657,584	265,197,659

Table 1: the sources and scales of the corpus.

It's a heavy task to manually classify those document into domains. However, we still can get the domain information for a certain subsets of the corpus. For some web sites listed above, we can get the domain information from the URL of each web page. For instance, the URL ”http://tb.chinatibetnews.com/xzmeishi/2011-12/05/content_831210.htm” shows it belongs to a column called ”xzmeishi”. so it must be a page about Tibetan foods, because ”xz” is the abbreviated form of Chinese word ”xizang” (西藏), which means the Tibetan Autonomous Region, while ”meishi” means ”delicious food”. So we can classify the documents in the corpus into domains. Table 2 and 3 list the domains of subsets of the documents from two web sites named ”China Tibet News” and ”Tibetan's web of China” respectively. Obviously, a large part of the documents in the corpus are news as expected, because nearly all of the 8 web sites are hold by news agencies or radio stations.

4.2 Methods to Segment Tibetan Text

As described in section 2, there is a delimiter between two Tibetan syllables. So we can segment the text into syllables by adding segmentation mark after the delimiter. The encoding schema can be used to segment text into smaller language units.

The challenge lies in the word segmentation. With a similar method to those methods proposed by other researchers (Chen et al., 2003a; Chen et al., 2003b; Jiang and Kong, 2006; Sun et al., 2009; Sun et al., 2010; Liu et al., 2012a), we implemented a segmenter. As mentioned in Section 2.2, some monosyllable words can glue to the previous word without a syllable delimiter “tsheg”, which produce many

Order	Domain	♯document	(%)	♯sentence	(%)	♯syllable	(%)
1	Art	3,240	4.76	112,642	8.71	1,265,914	4.40
2	Finance & Economy	712	1.05	12,477	0.96	314,698	1.09
3	History & Geometry	2,897	4.25	19,627	1.52	283,621	0.98
4	News	25,247	37.08	576,842	44.59	14,753,178	51.23
5	Picture	12,732	18.70	51,088	3.95	766,895	2.66
6	Politics & Law	3,230	4.74	63,437	4.90	1,708,839	5.93
7	Rural Life	2,402	3.53	35,535	2.75	871,406	3.03
8	Social Life	1,153	1.69	9,881	0.76	233,454	0.81
9	Special Issues	9,986	14.67	268,003	20.72	6,499,488	22.57
10	Technology & Education	1,988	2.92	38,321	2.96	825,395	2.87
11	Tibetan Buddhism	1,983	2.91	48,832	3.77	569,756	1.98
12	Tibetan Food	215	0.32	2,963	0.23	35,365	0.12
13	Tibetan Medicine	720	1.06	36,676	2.84	303,012	1.05
14	Tour	1,588	2.33	17,296	1.34	367,226	1.28
15	Total	68,093	100.00	1,293,620	100.00	28,798,247	100.00
Total		68,093	100.00	1,293,620	100.00	28,798,247	100.00

Table 2: Domains of a subset of the documents from ”China Tibet News”.

Order	Domain	♯document	(%)	♯sentence	(%)	♯syllable	(%)
1	Art	92	0.35	3,021	0.45	44,727	0.43
2	Culture	885	3.40	109,749	16.18	980,554	9.32
3	Economy	78	0.30	7,749	1.14	124,101	1.18
4	Education	15	0.06	695	0.10	13,919	0.13
5	Music	323	1.24	3,169	0.47	31,791	0.30
6	News	24,055	92.45	519,576	76.61	8,783,626	83.50
7	Photo	80	0.31	2,548	0.38	35,982	0.34
8	Policy	116	0.45	7,062	1.04	121,930	1.16
9	Politics	124	0.48	7,668	1.13	137,538	1.31
10	Tibetan Medicine	107	0.41	11,417	1.68	162,557	1.55
11	Tour	145	0.56	5,563	0.82	82,443	0.78
Total		26,020	100.00	678,217	100.00	10,519,168	100.00

Table 3: Domains of a subset of the documents from ”Tibetan’s web of China”.

abbreviated syllables. So Tibetan has a significant number of complex words where the sounds have been synthesized due to internal sandhi something like Sanskrit. As some of those abbreviated syllables can also be used as normal syllables, they lead to considerable problem in Tibetan word segmentation. So in the first step, we analyse the structure of each syllable in the sentence, and break them into normal syllables and abbreviated mark candidates and take them as the basic units (unbreakable units). Then, in the second step, some special case-auxiliary words (which are all monosyllable words) are used as separators to break the sentence into blocks. Consequently, both the forward maximum matching method and backward maximum matching method are used to segment each block into words. Mean while, it detects ambiguities by bidirectional segmentation, and makes disambiguation with word frequency. A previous research shows that the precision of this method reaches 96.98% (Liu et al., 2012a). The following example shows the main procedure of the method.

Input: ངཚོས་སྤྱི་ཚོགས་རིང་ལུགས་ཀྱི་སྤྱི་ལ་དབང་བའི་ལམ་ལུགས་དང་རྩོལ་བསྡུན་ཐོབ་སྤྲོད་ཀྱི་རྩ་དོན་མཐའ་འཁྱོངས་བྱས་ཡོད།

Translation: We have always followed the principles of socialist public ownership and distribution according to work.

Step 1: ང (ཚོ ས་) སྤྱི་ ཚོགས་ རིང་ ལུགས་ ཀྱི་ སྤྱི་ ལ་ དབང་ (བ འི་) ལམ་ ལུགས་ དང་ རྩོལ་ བསྡུན་ ཐོབ་ སྤྲོད་ ཀྱི་ རྩ་ དོན་ མཐའ་ འཁྱོངས་ བྱས་ ཡོད །

Step 2: (ང ཚོ ས་ སྤྱི་ ཚོགས་ རིང་ ལུགས་) ཀྱི་ (སྤྱི་ ལ་ དབང་ བ འི་ ལམ་ ལུགས་ དང་ རྩོལ་ བསྡུན་ ཐོབ་ སྤྲོད་) ཀྱི་ (རྩ་ དོན་ མཐའ་ འཁྱོངས་ བྱས་ ཡོད) །

Step 3: (ང ཚོ) (ས་) (སྤྱི་ ཚོགས་ རིང་ ལུགས་) ཀྱི་ (སྤྱི་ ལ་ དབང་ བ འི་ ལམ་) (ལུགས་) (དང་) (རྩོལ་) (བསྡུན་) (ཐོབ་) (སྤྲོད་) ཀྱི་ (རྩ་ དོན་) (མཐའ་ འཁྱོངས་) (བྱས་ ཡོད) །

Output: ངཚོ ས་ སྤྱི་ཚོགས་རིང་ལུགས་ ཀྱི་ སྤྱི་ལ་དབང་བའི་ལམ་ ལུགས་ དང་ རྩོལ་ བསྡུན་ ཐོབ་ སྤྲོད་ ཀྱི་ རྩ་དོན་ མཐའ་འཁྱོངས་ བྱས་ཡོད །

Note that, in step 1, two abbreviated syllable candidates are found and given in parentheses. In step 2, the two occurrences of the case-auxiliary word ཀྱི་ break the sentence into several blocks, and each block is segmented into words consequently in step 3. In this paper, as we mainly focus on Tibetan text, so in the segmentation, all Latin words, Latin numbers, Chinese phrases, Tibetan alphabetic numbers such as " ༦༣༣༡༠༨༩ " (6331089) and so on are all replaced by place-holders.

4.3 Counting and Calculation

The SRI Language Modelling Toolkit (SRILM) (Stolcke and others, 2002) is used to count the frequencies in our work.

5 Statistical Data and Analysis

In this section, we show the statistical data and check whether the frequency-rank on the units of super character, syllable and word follows Zipf's law respectively. As there isn't many enough items in the frequency list, we won't check it on the units of letter and character. For the other units, the number of occurrences of each n-gram is listed in Table 4.

unit	super char	syllable	word
unigram	265,197,659	93,657,584	67,209,927
bigram	260,565,029	89,024,954	62,577,297
trigram	256,022,772	84,521,746	58,085,930
4-gram	251,638,774	80,194,075	54,051,877
5-gram	247,285,278	76,155,897	50,242,999
Total	1,280,709,512	423,554,256	292,168,030

Table 4: Number of occurrences of each Tibetan n-gram in different units in the corpus.

5.1 Letter Frequency

In total, Tibetan 83 letters are used in the corpus, of which 39 letters are consonants and 8 letters are vowel signs. Letter ཷ and ཹ didn't occur in the corpus, which means people might prefer to use two letters to spell each of them. The other are Tibetan punctuations and signs. There are also 200 non Tibetan characters used in the corpus. The 47 letters and the two delimiters are listed in Table 5. The character "P", "C" and "V" in the table denote "Punctuation", "Consonant" and "Vowel" respectively. The "theg" shares 23.45% of the corpus while the "thed" shares 1.33%, which shows that a syllable has 3.26 letters (not including the "theg" itself), while a sentence has 75.40 letters in average. 4 of the 8 vowels occur frequently while the other 4 vowels are rarely used. The 2 punctuations share 24.7762% of the corpus while 4 vowels in modern Tibetan share 16.0805%, and the 30 consonants in modern Tibetan share 58.3918%. All these 36 letters share 99.2485% of the corpus in total. The other 4 vowels and 9 consonants which is used to transliterate Sanskrit script are rarely used. They share only 0.0437%. Other Tibetan signs and non Tibetan characters share 0.7078%.

5.2 Character Frequency

There are 119 Tibetan characters used in the corpus in total, including Tibetan punctuations and signs, but Tibetan number is replaced with a place-holder. As there are 83 letters as described in the former subsection, the other 36 characters are the second or third forms of Tibetan consonants. As the frequency of Tibetan character is seldom a concerned issue, we don't make any further remarks on it.

5.3 Super Character Frequency

There are 1,466 super characters used in the corpus in total. The topmost frequently occurred super characters and n-gram super char phrases for $n = 2, 3$ are listed in Table 6. As expected, the "theg" is the most frequently occurred one when we take it as a super character, which shares 31.22%. It indicates that a syllable is formed by 2.20 super characters in average. The "theg" shares 1.5837%, which indicates that a sentence has 63.14 super characters in average.

♯	letter	♯ occur	rate(%)	cum.rate(%)	♯	letter	♯ occur	rate(%)	cum.rate(%)
P01	་	82,818,775	23.4500	23.4500	C20	ཅ	1,788,360	0.5064	96.3522
C01	ས	25,967,545	7.3527	30.8027	C21	ཕ	1,688,121	0.4780	96.8302
C02	ག	21,589,015	6.1129	36.9155	C22	ཉ	1,442,603	0.4085	97.2386
C03	ར	18,211,934	5.1567	42.0722	C23	ཙ	1,193,175	0.3378	97.5765
V01	ེ	17,755,843	5.0275	47.0998	C24	ཤ	1,110,743	0.3145	97.8910
V02	ི	17,663,843	5.0015	52.1012	C25	ཟ	1,106,366	0.3133	98.2043
C04	ད	16,796,414	4.7559	56.8571	C26	ཇ	1,074,769	0.3043	98.5086
C05	ང	14,320,083	4.0547	60.9118	C27	ཧ	962,770	0.2726	98.7812
C06	བ	14,311,260	4.0522	64.9640	C28	ཛ	932,081	0.2639	99.0451
C07	ཡ	13,969,128	3.9553	68.9194	C29	ཨ	411,338	0.1165	99.1616
V03	ུ	12,067,901	3.4170	72.3364	C30	ཝ	306,863	0.0869	99.2485
C08	ན	11,706,716	3.3147	75.6511	V05		112,528	0.0319	99.2803
C09	ལ	10,652,623	3.0163	78.6674	C31	ཎ	13,398	0.0038	99.2841
C10	འ	10,252,221	2.9029	81.5703	C32	ཥ	11,482	0.0033	99.2874
V04	ོ	9,304,303	2.6345	84.2048	C33	ཌ	7,979	0.0023	99.2896
C11	མ	8,003,478	2.2662	86.4709	C34	ཊ	6,617	0.0019	99.2915
C12	པ	6,484,634	1.8361	88.3070	C35	བྷ	945	0.0003	99.2918
C13	ཀ	4,788,018	1.3557	89.6628	V06	ྀ	689	0.0002	99.2920
P02	།	4,683,745	1.3262	90.9890	V07	ཽ	288	0.0001	99.2920
C14	ཁ	3,621,446	1.0254	92.0144	C36	གྷ	159	0.0000	99.2921
C15	ཏ	3,220,311	0.9118	92.9262	C37	དྷ	128	0.0000	99.2921
C16	ཚ	3,082,972	0.8729	93.7991	V08	ཻ	64	0.0000	99.2921
C17	ཞ	2,696,958	0.7636	94.5628	C38	ཀྵ	55	0.0000	99.2922
C18	ཐ	2,307,810	0.6535	95.2162	C39	ཬ	38	0.0000	99.2922
C19	ཆ	2,223,533	0.6296	95.8458	Total		353,171,951		100.00

Table 5: Frequency of Tibetan letters used in the corpus.

♯	Unigram	♯occur	rate(%)	Bigram	♯occur	rate(%)	Trigram	♯occur	rate(%)
1	་	82,794,773	31.2200	ས་	15,628,059	5.9978	གས་	3,507,130	1.3699
2	ས	17,136,507	6.4618	ང་	11,178,538	4.2901	དང་	2,344,655	0.9158
3	ང	13,283,592	5.0089	ན་	7,913,908	3.0372	་དང	2,311,335	0.9028
4	ག	12,677,652	4.7805	་བ	6,137,354	2.3554	་པ་	1,829,894	0.7147
5	ད	11,999,418	4.5247	ད་	6,108,586	2.3444	ངས་	1,699,226	0.6637
6	ན	10,582,545	3.9904	་འ	4,972,884	1.9085	་ཀྱི་	1,353,391	0.5286
7	བ	9,387,744	3.5399	ར་	4,892,373	1.8776	་པའི	1,284,801	0.5018
8	ལ	6,335,567	2.3890	་ད	4,705,897	1.8060	པའི་	1,282,074	0.5008
9	མ	5,899,668	2.2246	ལ་	4,596,516	1.7641	ནས་	1,251,081	0.4887
10	འ	5,769,971	2.1757	་ག	4,268,149	1.6380	ས་པ	1,244,612	0.4861
Total		175,867,437	66.3156	Total	70,402,264	27.0191	Total	18,108,199	7.0729

Table 6: The topmost frequently occurred super characters and n-gram super char phrases.

The frequency-rank curves of the n-gram for $n = 1, 2, 3, 4, 5$ are plotted with log-log axes in Figure 5. A straight line with $slop = -1.0$ is also plotted in the figure. It's obvious that the curves don't follow Zipf's law so exactly. The high frequency parts of the curves follow Zipf's law at large, but as the rank increases the curves have more rapid decreases than a linear curve with slop $= -1.0$ when the rank > 100. However, we still found that the curve becomes more straight when the n increases.

Similar to Ha et al. (2002), we also combine the frequency list of the n-grams for all $n = 1, 2, \ldots, 5$ together in one list and put in order of frequency. The frequency-rank curve is plotted in Figure 6. A straight line with slope $= -1.0$ is also plotted in the figure, which shows that there are large gaps between

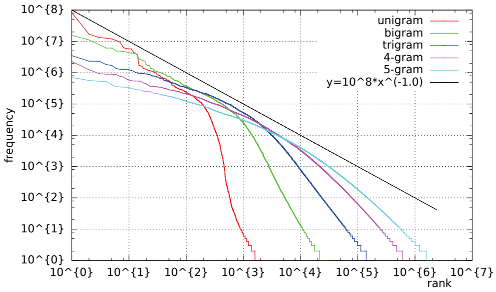


Figure 5: Frequency-rank of super chars and their n-grams.

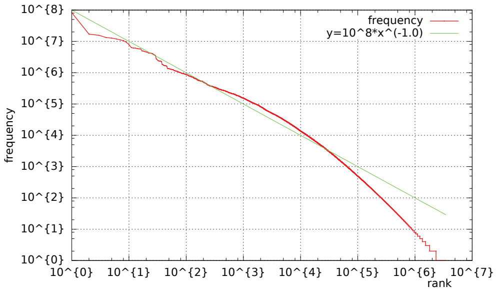


Figure 6: Frequency-rank of combined super char n-gram list.

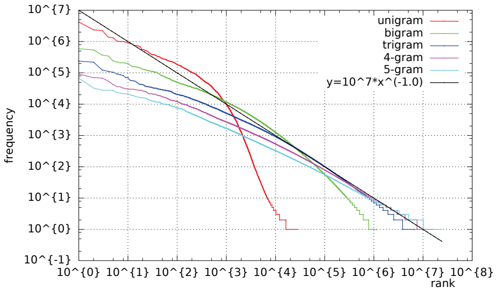


Figure 7: Frequency-rank of syllables and syllable n-grams.

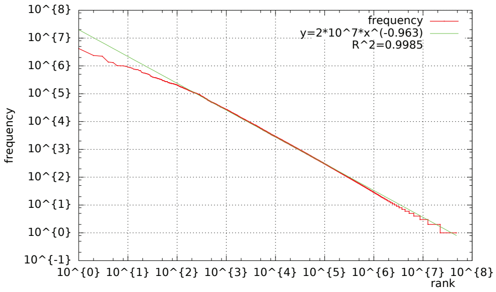


Figure 8: Frequency-rank of combined syllable n-gram list.

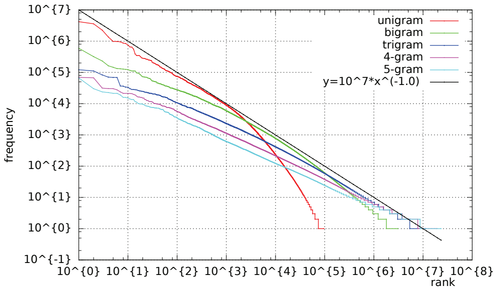


Figure 9: Frequency-rank of words and word n-grams.

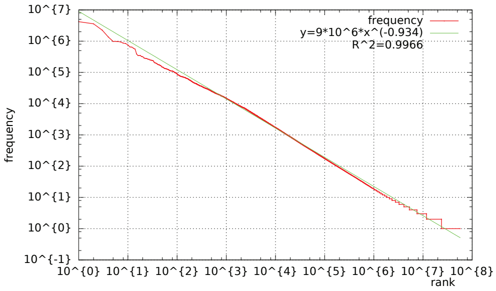


Figure 10: Frequency-rank of combined word n-gram list.

the curve and the line. Obviously it doesn't follow Zipf's law well.

5.4 Syllable Frequency

There are 27,546 syllables and 200 other characters occurred in the corpus in total. The topmost frequently occurred syllables and n-gram syllable phrases for $n = 2, 3$ are listed in Table 7. As expected, the "thed" is the most frequently used unigram when we take is as a syllable. It shares 4.4843% of the corpus. Most of the top 15 unigrams are case auxiliary words (monosyllable word), including གྱི་ , ལ་ , ལས་ , ནས་ , ཀྱི་ and གི་ . The conjunction དང་ , the two nominalization markers པ་ and བ་ are also in the top 10 list. The top 10 syllables take up 16.5556% of the corpus.

The frequency-rank curves of the n-gram for $n = 1, 2, 3, 4, 5$ are plotted with log-log axes in Figure 7. A straight line with $slop = -1.0$ is also plotted in the figure , which shows that the curves don't follow Zipf's law very exactly. The high frequency parts of the curves when $n = 1, 2$ follow Zipf's law at large, but as the rank increases the curves have more rapid decreases than a linear curve with slop $= -1.0$ when the rank > 1000 and the rank > 10000 respectively. The curve becomes more straight when the n increases, and becomes almost straight lines when $n = 3, 4, 5$.

We also combine the frequency list of the n-grams for all $n = 1, 2, \ldots, 5$ together in one list and

♯	Unigram	♯occur	rate(%)	Bigram	♯occur	rate(%)	Trigram	♯occur	rate(%)
1	।	4,199,896	4.4843	དང་ ।	593,668	0.6669	པ་ དང་ ।	237,068	0.2805
2	དང་	2,370,981	2.5315	པ་ དང་	537,089	0.6033	པ་ རེད་ ।	219,016	0.2591
3	པ་	2,233,002	2.3842	པ་ ।	367,725	0.4131	རང་ སྐྱོང་ ལྗོངས་	104,505	0.1236
4	ཀྱི་	1,377,206	1.4705	རེད་ ।	345,059	0.3876	ཡོད་ པ་ རེད་	92,475	0.1094
5	པའི་	1,319,052	1.4084	ཡོད་ ।	241,152	0.2709	བྱེད་ པ་ དང་	86,964	0.1029
6	ལ་	1,023,287	1.0926	པ་ རེད་	222,771	0.2502	བ་ དང་ ।	82,532	0.0976
7	བ་	1,008,926	1.0772	བ་ དང་	209,657	0.2355	བྱེད་ པ་ ।	58,133	0.0688
8	ལས་	1,007,539	1.0758	ཡོད་ པ་	209,095	0.2349	ཡོད་ པ་ དང་	56,153	0.0664
9	ནས་	965,728	1.0311	བྱེད་ པ་	205,235	0.2305	ཨུ་ ཡོན་ ལྷན་	55,173	0.0653
10	བྱེད	915,663	0.9777	བ་ ।	198,682	0.2232	ཡོན་ ལྷན་ ཁང་	54,168	0.0641
Total		15,505,617	16.5556	Total	2,931,451	3.2928	Total	992,019	1.1737

Table 7: The topmost frequently occurred syllables and n-gram syllable phrases.

put in order of frequency. The frequency-rank curve is plotted in Figure 8. A fitting straight line $y = 2 \times 10^7 \times x^{-0.963}$ with $R^2 = 0.9985$ is also plotted in the figure, which shows that the curve can be well fitted by the line. Thus, it follows Zipf's law.

5.5 Word Frequency

♯	Unigram	♯occur	rate(%)	Bigram	♯occur	rate(%)	Trigram	♯occur	rate(%)
1	।	4,199,896	6.2489	དང་ ।	593,286	0.8258	ཡོད་པ་ རེད་ ।	83,110	0.1237
2	འི་	3,580,891	5.3279	རེད་ ।	319,479	0.4447	ལོ་ འི་ ཟླ་	36,087	0.0537
3	དང་	2,241,125	3.3345	ཡོད་ ।	232,155	0.3231	བྱེད་པ་ དང་ ।	33,574	0.0500
4	ཀྱི་	1,357,874	2.0203	པ་ འི་	163,093	0.2270	ཡོད་པ་ དང་ ।	28,562	0.0425
5	ལ་	982,161	1.4613	༄༅ ।།	145,563	0.2026	བྱས་ ཏེ་ ।	28,387	0.0422
6	ར་	977,484	1.4544	བྱེད་པ་ འི་	132,517	0.1845	ཚོམ་སྒྲིག་ འགན་ཁུར་བ་ ।	27,707	0.0412
7	ནས་	949,750	1.4131	བོད་ ཀྱི་	131,176	0.1826	གཞིགས་ ན་ ।	26,523	0.0395
8	གྱི་	863,061	1.2841	བྱས་ ཏེ་	116,202	0.1617	བྱས་པ་ རེད་ ।	25,829	0.0384
9	གི་	754,281	1.1223	བྱས་ ནས་	110,635	0.1540	ལྗོངས་ ཡོངས་ ཀྱི་	25,418	0.0378
10	ས་	654,987	0.9745	ཀྲུང་གོ་ འི་	96,640	0.1345	བྱས་ ཡོད་ ।	24,786	0.0369
Total		16,561,510	24.6415	Total	2,040,746	2.8406	Total	339,983	0.5058

Table 8: The topmost frequently occurred words and n-gram word phrases.

There are 96,296 words(including Tibetan punctuations, signs) used in the corpus in total. The topmost frequently occurred words and n-gram word phrases for $n = 2, 3$ are listed in Table 8. As expected, the "thed" is the most frequently used unigram when we take is as a word. It shares 6.2489% of the corpus. Almost all of the top 10 unigrams are auxiliary case words (monosyllable word), including འི་ , ཀྱི་ , ལ་ , ར་ , ནས་ , གྱི་ , གི་ and ས་ . The top 10 words take up 24.6415% of the corpus.

The frequency-rank curves of the n-gram for $n = 1, 2, 3, 4, 5$ are plotted with log-log axes in Figure 9. A straight line with $slop = -1.0$ is also plotted in the figure , which shows that the curves don't follow Zipf's law very exactly. The high frequency part of the curve when $n = 1$ follows Zipf's law at large, but as the rank increases the curve has more rapid decreases than a linear curve with slop $= -1.0$ when the rank > 1000. The curve becomes more straight when the n increases, and becomes almost straight lines when $n = 3, 4, 5$.

We also combine the frequency list of the n-grams for all $n = 1, 2, \ldots, 5$ together in one list and put in order of frequency. The frequency-rank curve is plotted in Figure 10. A fitting straight line

$y = 2 \times 10^7 \times x^{-0.934}$ with $R^2 = 0.9966$ is also plotted in the figure, which shows that the curve can be well fitted by the line. Thus, it follows Zipf's law.

5.6 Further Discussion

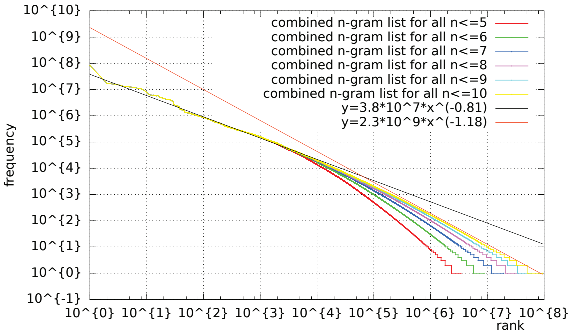


Figure 11: Frequency-rank of combined Tibetan word n-gram lists for $n <= 5, 6, 7, 8, 9, 10$.

Comparing the curves in Figure 5, 7 and 9, we find that the curves with the same n for all $n = 1, 2, 3, 4, 5$ become more straight when the granularity becomes larger. It's similar in the combined n-gram curves in Figure 6, 8 and 10. As it's shown that the two combined n-gram frequency lists for all $n <= 5$ on syllable and word follow Zipf's law well. So, the question is that whether we can find a larger M, which for the combined n-gram list for all all $n < m$, the frequency-rank curve in log-log axes is straight enough. To find the M, the frequency-rank curves for the combined n-gram super character lists for $m = 5, 6, 7, 8, 9, 10$ are plotted respectively in Figure 11. From the figure, we see that the head parts of the curves are overlapped, which correspond to the high frequency parts of the combined n-gram lists, while the tail parts of the curves are divergent. As the m increases, the tail part of the curve becomes closer to the straight line $y = 3.8 \times 10^7 \times x^{-0.81}$. This mainly results from that the frequency of the n-gram decreases when the n increase, and the low frequency part of the combined n-gram list includes more n-grams. However, the two straight lines of $y = 3.8 \times 10^7 \times x^{-0.81}$ and $y = 2.3 \times 10^9 \times x^{-1.18}$ in the figure show that any one of those curves can't be fitted well by a straight line. The reason leading to this somewhat unusual result is an issue to be made further research and analysis.

6 Conclusion

In the former section, we make statistics on different Tibetan language units : letter, super character, syllable and word, and their n-gram phrases. It shows that when we put all the n-gram phrases for $n = 1, 2, \ldots, 5$ together and sort all of them by frequency in descending order, then the frequency-rank curves in log-log axes can be fitted well for the unit of syllable and word respectively. But for the unit of super character, we didn't find a curve which can be fitted well enough by a straight line when we combine all the n-grams for $n <= m$ even if m is up to 10.

Acknowledgements

We thank the reviewers for their critical and constructive comments and suggestions that helped us improve the quality of the paper. The research is partially supported by National Science and Technology Major Project (No.2012ZX01039-004), National Science Foundation (No.61202219, No.61202220, No.61303165), Major Science and Technology Projects in Press and Publishing (No.0610-1041BJNF2328/23, No.0610-1041BJNF2328/26), and Informationization Project of the Chinese Academy of Sciences (No.XXH12504-1-10).

References

Lada A. Adamic and Bernardo A. Huberman. 2002. Zipfs law and the internet. *Glottometrics*, 3(1):143–150.

Jinyong Ai, Hongzhi Yu, and Yonghong Li. 2009. Statistical analysis on tibetan shaped structure. *Journal of Computer Applications*, 29(7):2029–2031.

MG Boroda and AA Polikarpov. 1988. The zipf-mandelbrot law and units of different text levels. *Musikometrika*, 1:127–158.

Lee Breslau, Pei Cao, Li Fan, Graham Phillips, and Scott Shenker. 1998. On the implications of zipfs law for web caching. Technical report, Citeseer.

CA Chatzidimitriou-Dreismann, RMF Streffer, and Dan Larhammar. 1996. Lack of biological significance in the linguistic features of noncoding dnaa quantitative analysis. *Nucleic acids research*, 24(9):1676–1681.

Yuzhong Chen, Baoli Li, and Shiwen Yu. 2003a. The design and implementation of a tibetan word segmentation system. *Journal of Chinese Information Processing*, 17(3):15–20.

Yuzhong Chen, Baoli Li, Shiwen Yu, and Lancuoji. 2003b. An automatic tibetan segmentation scheme based on case auxiliary words and continuous features. *Applied Linguistics*, 2003(01):75–82.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

The Unicode Consortium. 2013. *The Unicode Standard, Version 6.3.0*. The Unicode Consortium, ISBN 978-1-936213-08-5, Mountain View, CA.

Leo Egghe. 1999. On the law of zipf-mandelbrot for multi-world phrases.

Leo Egghe. 2000. The distribution of n-grams. *Scientometrics*, 47(2):237–252.

Xavier Gabaix. 1999a. Zipf's law and the growth of cities. *The American Economic Review*, 89(2):129–132.

Xavier Gabaix. 1999b. Zipf's law for cities: an explanation. *The Quarterly Journal of Economics*, 114(3):739–767.

Dingguo Gao and Yuchang Gong. 2005. A statistically study on the qualities of all modern tibetan character set. *Journal of Chinese Information Processing*, 19(1):71–75.

Le Quan Ha, Elvira I Sicilia-Garcia, Ji Ming, and F Jack Smith. 2002. Extension of zipf's law to words and phrases. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–6. Association for Computational Linguistics.

Yannis M Ioannides and Henry G Overman. 2003. Zipfs law for cities: an empirical examination. *Regional science and urban economics*, 33(2):127–137.

Di Jiang and Yinghong Dong. 1994. Statistical analysis on linear processing of tibetan clustered structures. *Chinese Information Processing*, (4):44–46.

Di Jiang and Yinghong Dong. 1995. Research on property of tibetan characters as information processing. *Journal of Chinese Information Processing*, 9(2):37–44.

Di Jiang and Jiangping Kong. 2006. *Advances on the Minority Language Processing of China*. Social Sciences Academic Press, Beijing, China.

Di Jiang and Congjun Long. 2010. *On Characters of Tibetan Writing System: Alpabetic Characters, Pronunciations, ISO Codes, Frequencies, Sorting Orders, Picture Symbols and Transliterations*. Social Sciences Academic Press, Beijing, China.

Di Jiang. 1998. An entropy value of classical tibetan language and some other questions. In *Proceedings of International Conference on Chinese Information Processing*, pages 377–381. Chinese Information Processing Society of China.

Di Jiang. 2006. History and development of tibetan text information processing. In *Frontiers of Chinese Information Processing - Proceedings of the 25th Anniversary Conference of Chinese Information Processing Society of China*, pages 83–97. Tsinghua University Press, Beijing, China.

Wentian Li. 1992. Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6):1842–1845.

Huidan Liu, Minghua Nuo, Longlong Ma, and et al. 2011. Tibetan Word Segmentation as Syllable Tagging Using Conditional Random Fields. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 2011)*, pages 168–177.

Huidan Liu, Minghua Nuo, Longlong Ma, and et al. 2012a. SegT: A pragmatic tibetan word segmentation system. *Journal of Chinese Information Processing*, 26(1):97–103.

Huidan Liu, Minghua Nuo, Jian Wu, and Yeping He. 2012b. Building large scale text corpus for tibetan by extracting text from web pages. In *Proceedings of the 10th asian language resources at COLING 2012*, pages 8–17.

Yajun Lu and Xiaodong Shi. 2011. Random texts exhibit zipf's-law-like word frequency distribution. *Journal of Chinese Information Processing*, 25(4):54–56.

Yajun Lu, Shaoping Ma, Min Zhang, and Guang Luo. 2003. Researches of calculations of tibetan characters, pieces, syllables, vocabulary and universal frequency and its applications. *Journal of Northwest Minorities University(Natural Science)*, 24(48):32–42.

R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, and H. E. Stanley. 1994. Linguistic features of noncoding dna sequences. *Phys. Rev. Lett.*, 73:3169–3172, Dec.

AIDA Masaki and Noriyuki Takahashi. 1998. A proposal of dual zipfian model for describing http access trends and its application to address cache design. *IEICE transactions on communications*, 81(7):1475–1485.

Marcelo A Montemurro and Damian H Zanette. 2002. New perspectives on zipfs law in linguistics: from single texts to large corpora. *Glottometrics*, 4:87–99.

S Naranan and VK Balasubrahmanyan. 1998. Models for power law relations in linguistics and information science. *Journal of Quantitative Linguistics*, 5(1-2):35–61.

Kazumi Okuyama, Misako Takayasu, and Hideki Takayasu. 1999. Zipf's law in income distribution of companies. *Physica A: Statistical Mechanics and its Applications*, 269(1):125–131.

Ronald Rousseau. 2002. George kingsley zipf: life, ideas, his law and informetrics. *Glottometrics*, 3:11–18.

Herbert A Simon. 1955. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440.

Kwok Tong Soo. 2005. Zipf's law for cities: a cross-country investigation. *Regional science and urban Economics*, 35(3):239–263.

Kwok Tong Soo. 2007. Zipf's law and urban growth in malaysia. *Urban Studies*, 44(1):1–14.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 901–904. Denver.

Yuan Sun, Luosangqiangba, Rui Yang, and Xiaobing Zhao. 2009. Design of a tibetan automatic segmentation scheme. In *the 12th Symposium on Chinese Minority Information Processing*.

Yuan Sun, Xiaodong Yan, , Xiaobing Zhao, and Guosheng Yang. 2010. A resolution of overlapping ambiguity in tibetan word segmentation. In *Proceedings of the 3rd International Conference on Computer Science and Information Technology*, pages 222–225.

Weilan Wang and Wanjun Chen. 2004. The frequency and information entropy of tibetan character and syllabie. *Terminology Standardization and Information Technology*, (2):27–31.

Weilan Wang. 2004. The frequency-rank of language unit in modern tibetan. *Science Technology and Engineering*, 4(5):413–417.

Damián Zanette and Marcelo Montemurro. 2005. Dynamics of text generation with realistic zipf's distribution. *Journal of quantitative Linguistics*, 12(1):29–40.

Damiáan H Zanette. 2006. Zipf's law and the creation of musical context. *Musicae Scientiae*, 10(1):3–18.

George Kingsley Zipf. 1935. The psycho-biology of language.

George Kingsley Zipf. 1949. Human behavior and the principle of least effort.