IOP Conf. Series: Materials Science and Engineering **317** (2018) 012057 doi:10.1088/1757-899X/317/1/012057

Zipf's Law and the Frequency of Kazak Phonemes in Word Formation

Ruiqing Xin[1], Yonghong Li[1] and Hongzhi Yu[1]

[1]Key Lab of China's National Linguistic Information Technology, Northwest Minzu University, Lanzhou, 730030, China

Abstract. Zipf's Law is the basis of the principle of Least Effort, and is widely applicable in all natural fields. The occurring frequency of each phoneme in all Kazak words has been counted to testify the application of Zipf's law in Kazak. Due to the limitation of the sample size, deviation is unavoidable, but overall results indicate that the occurring frequency and the reciprocal rank of each phoneme in Kazak words formation are in line with Zipf's distribution.

1. Introduction

Modern Kazak is a common language for all Kazak people in the world. At present there are mainly two writing systems of Kazak. One is in Cyrillic for Kazakhstan, the other one is in Arabic for China. Although the writing systems are different, the pronunciations of the phonemes are almost the same, and all Kazak people can communicate well. In China, there are altogether 33 phonemes in Kazak, among which 9 are vowels and 24 are consonants [1]. The phonetic system of Kazak in China is shown in table 1.

Table 1. Kazak phoneme system

	Kazak	IPA		Kazak	IPA
Front vowels	ە	e	**Voiced consonants**	ب	b
	ٵ	æ		د	d
	ٶ	ø		گ	g
	ٷ	y		ق	q
	ٸ	i		ج	ʤ
Back vowels	ا	ɑ		ۋ	v
	و	o		ز	z
	ۇ	u		ع	ʁ
	ى	ə		م	m
Voiceless consonant	پ	p		ن	n
	ت	t		ل	l
	ك	k		ڭ	ŋ
	چ	ʧ		ر	r
	ف	f	**Approximant**	ي	j
	س	s		ۉ	w
	ش	ʃ			
	ح	x			
	ھ	h			

Great progress has been made in the research in Kazak phonetics. *Modern Kazak* (Kazak version) written by Geng Shimin & Li Zengxiang and *Modern Kazak* (Kazak version) composed by the Language Committee of Xinjiang Province both make a complete description of Kazak phonetics including the biological and physical properties of the phonemes, the principles of phonetic harmony, the phenomena of phonetic changes in speech, so on and so forth; while *Kazak Grammar* compiled by the Institute of Linguistics of the Education and Science Department of the Republic of Kazakhstan offers the most detailed depiction of Kazak phonetics [2]. But in the formation of words, what functions does each phoneme play? What is the occurring frequency of each phoneme? These are basic but interesting topics. Zipf's Law will be taken to analyze the occurring frequency of each phoneme in all Kazak words.

2. Zipf's Law

The principle, which means to get the greatest achievements with the least effort, has been known as the principle of Least Effort, or the Economy Principle, and it has been taken as one primary rule to monitor human's behavior. The principle was firstly raised by George Kinsley Zipf, based on what is called Zip's Law.

Zipf found that in a long discourse, the rank of a word (simplified as "r") which indicates the ranking of the occurring frequency of a word in the discourse, and the occurring frequency of the word (simplified as "f") have a certain relationship. He explained the relationship with the equation "r×f=c", in which "c" is "constant". That is to say, the rank of a word and its occurring frequency in the discourse are in inverse ratio. The higher the occurring frequency is, the smaller the rank is. The word which occurs most frequently in the discourse is ranked in the first place.

The most attractive is that the result of "r×f" is a constant. Zipf analysed the words in *Ulysses*, and got the amazing finding—the product of the rank and occurring frequency of each word is almost the same, that is 26, 000 approximate, and he draw the doubly logarithmic chart according to the research results, which is shown in figure 1[3].

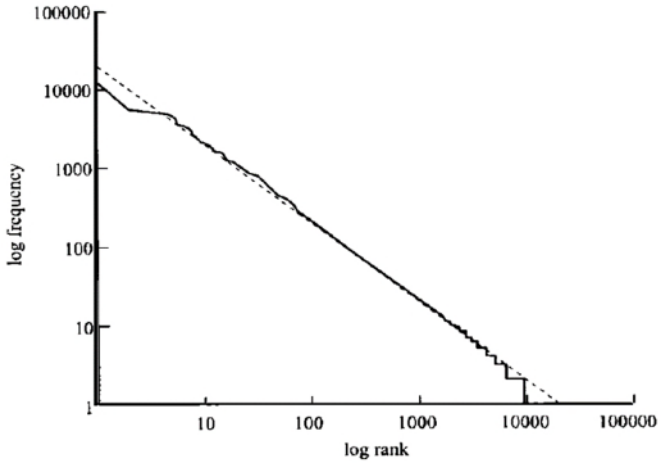

Figure 1. The Doubly Logarithmic Chart of Words in *Ulysses*.

In the chart, the line associating the frequency and the rank of the words is almost a straight line of 45 degrees. Of course, the result is just an average assessment. The relation between "r" and "f" is known as Zipf's Law.

Starting from Zipf's Law, and with further researches, Zipf raised the principle of least effort. The understanding could be that the writer is always trying to express the most with the fewest words, and thus, some words which are more favored by the writer occur more frequently in the discourse.

3. Research design

Since Zipf's Law has been reported to be applicable in many fields, such as information science, computer science, economy, so on and so forth. Whether it is also applicable to phonemes in word formation of Kazak is the focus of this paper.

This research is possible thanks to the Internet Kazak Corpus—Kazak-Chinese Dictionary uploaded by Li Jing [4] and *Kazak-Chinese Dictionary* complied by Nurbek Abken [5]. To guarantee the validity of the research, each entry of word in the corpus was examined according to *Kazak-Chinese Dictionary*. A new corpus—Kazak Phoneme Corpus was set up with all the Kazak words elicited from the Internet Kazak Corpus—Kazak-Chinese Dictionary. Because in Kazak words with front vowels, there is a front vowel mark "ء" before the words, while the symbols of the vowels in the words are the same with the back vowels counterparts, e.g. "ءشىم", it is complicated to count the frequency of all the vowels. Thus, all phonemes were turned into IPA before the division, and every word was divided into phonemes with the function of "LEFT( )" in Excel 2007. Entries with different definitions but the same spellings were screened and omission was done to excessive words with the same spelling, in order to guarantee the precision of the phoneme frequency in word formation. The sample of the new corpus is shown in figure 2. The number of all the words in the corpus is 49622.

WN	IPA	KAZAK	Clas	Chinese Definition	FV	P1	P2	P3	P4	P5	P6
N48801	ilw Ⅰ	Ⅰ ءىلۋ	noun.	<名>（亲家之间的）礼物，礼品，赠品 ءىلۋ الۋ 接受礼物 | ءىلۋ بەرۋ 赠送礼品。	ء	i	l	w		Ⅰ	
N48803	əlwde	ىلۋدە	adj.	ءىبىر ىلۋدە 少有的，稀罕的，罕见的 ءىبىر كەزدەسەتىن نارسە ىلۋدە 罕见之物。		ə	l	w	d	e	
N48804	ilwli	ءىلۋلى	adj.	<形> 挂着的，悬挂着的；吊着的 ءىلۋلى كيىم 挂着的衣服。	ء	i	l	w	l	i	
N48805	ələgw	ىلىگۋ	verb1.	ىلىك Ⅱ 的不定式。		ə	l	ə	g	w	
N48806	ələgəs	ىلىگىس	verb.	<动> 争吵，闹别扭 ول مەنەن ىلىگىسىپ قالدى 他俩争吵起来了。		ə	l	ə	g	ə	s
N48807	əlgəsw	ىلگىسۋ	verb1.	ىلىگىس 的不定式。		ə	l	g	ə	s	w

※ WN=word number, FV= front vowel mark, P=phoneme place

Figure 2. Sample of Kazak Corpus.

4. Results analysis

4.1. Distribution of Kazak phonemes in word formation

With the function of "COUNTIF ( )", the occurring frequency of each phoneme was counted and then the phonemes were ranked according to the frequency. The result is shown in table 2.

Table 2. The frequency and rank of Kazak phonemes.

Kazak	Phoneme	Frequency	Rank	Kazak	Phoneme	Frequency	Rank
ا	a	49585	1	ۇ	u	6812	18
ى	ə	38698	2	ز	z	6601	19
ت	t	25058	3	ج	ʤ	5951	20
ل	l	21784	4	پ	p	5465	21
ه	e	20972	5	ڭ	ŋ	4028	22
ر	r	19520	6	ع	ʁ	3616	23
ق	q	17969	7	گ	g	2579	24
س	s	16129	8	ءى	i	2210	25
ن	n	13049	9	ۇ	æ	1289	26
د	d	12962	10	ۋ	y	617	27
ۋ	w	11948	11	ح	x	561	28
ي	j	11622	12	ف	f	445	29

ك	k	10951	13	ۇ	ø	397	30
ش	ʃ	9819	14	ۋ	v	241	31
و	o	8955	15	ھ	h	106	32
م	m	8971	16	ڃ	ʧ	37	33
ب	b	7802	17				

Based on the data in table 2, a curve of Kazak phonemes' frequency was drawn in figure 3. Although the curve in figure 3 is not the same as that in figure 1, the general tendency of the distribution is in a triangle with an angle of about 45 degrees, and with the help of extension lines the most steady part of the curve can be a line with an angle of about 45 degree.

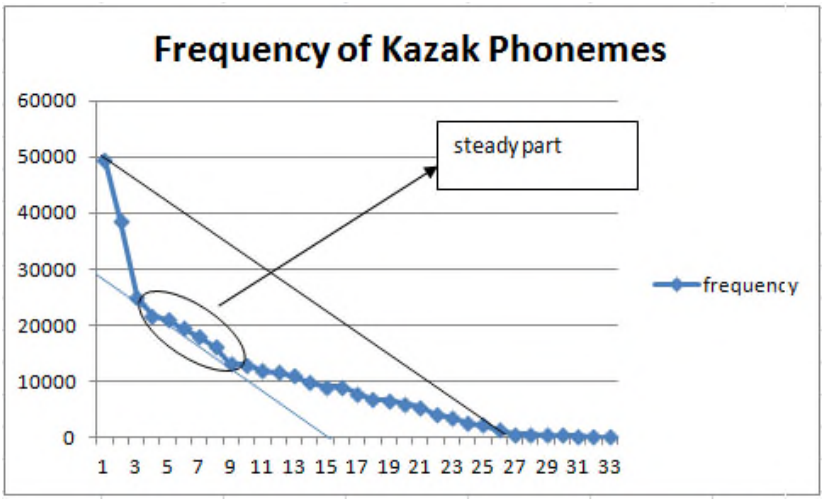

Figure 3. Curve of Kazak phonemes' frequency.

4.2. Application of Zipf's Law in Kazak phonemes' distribution

In figure 4 is the product curve of frequency and rank. According to the equation of Zipf's law "r×f=c", the average product of frequency and rank is a constant. Therefore, the product curve of Kazak phonemes' frequency and rank should almost be a line, which has been testified in figure 5, the logarithmic curve of the product of frequency and rank.

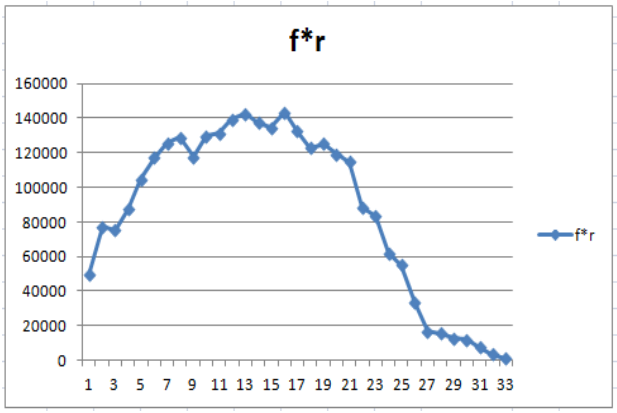

Figure 4. Product curve of frequency and rank.

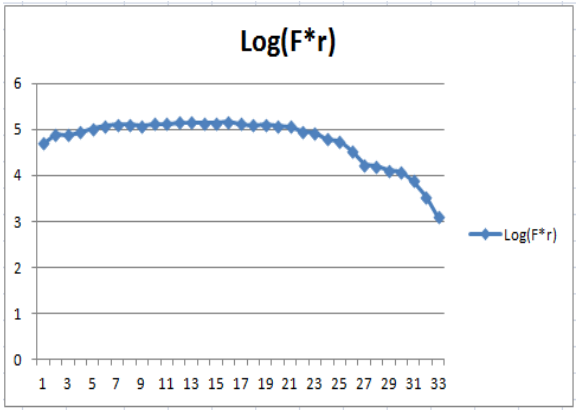

Figure 5. Logarithmic curve of the product of frequency and rank.

Data in table 2 and the curve in figure 3 have almost proved the application of Zipf's law in Kazak phonemes' distribution. The result is not exactly the same as Zipf's study of *Ulysess*, one of the possible reasons is that the sample of *Ulysess* is large, and the power of balance is great. Besides, the

frequency and the rank are in inverse ratio, so as a matter of fact, in any sample, both the variables ranked at the first places and those ranked at the last places should not be taken as examples for the law, because according to the equation "r×f=c", there will be two extremes in a sample—the sample is filled with the variables occurring most frequently and without any variables occurring least frequently. It is believed this is why not every variable was taken into consideration in the research of *Ulysess*. This explains the reason why only the most steady part of the curve, which is driven from the average data, is reliable in this research.

5. Conclusion

Zipf distribution is the result of balance and average, the two extremes of any distribution frequency in rank will not meet the law, which is why only the steady parts of the above curves are taken for study. This research has proved Zipf distribution is working in Kazak phonemes' occurring frequency in word formation, while more researches are expected.

References

[1] Geng S M & Li Z X 1985 *An Introduction to Kazak*. Beijing: The Ethnic Publishing House.
[2] Zhang D J 2009 *J. Yili Normal University (Soc. Sci. Ed.)*, **3** 25-39.
[3] Jiang W Q 2005 *J. Tongji University (Soc. Sci. Ed.)* **16(1)** 87-95.
[4] Li J 2015 Internet Kazak Corpus — Kazak-Chinese dictionary. http://vdisk.weibo.com/s/taoCn8PUHOqk4.
[5] Nurbek A 2014 *Kazak-Chinese Dictionary*. Beijing: The Ethnic Publishing House.

Acknowledgements

This paper is based on part of the research findings of Northwest Minzu University Graduate Innovation Project *The Acoustic Properties of Kazak Consonants* (Project No. YXM2016002).