UDC 81-139

G. Altenbek[1–3], X.L. Wang[1]

[1]*School of Computer Science and Technology, Harbin Institute of Technology, China;*
[2]*College of Information Science and Engineering, Xinjiang University, China;*
[3]*The Base of Kazakh and Kirghiz Language of National Language Resource Monitoring and Research Minority Language Centre, China*
(E-mail: gla@insun.hit.edu.cn)

A corpus-based frequency statistic of kazakh language

Kazakh language is an agglutinative language and it belongs to the Turkish Language group. Kazakh is a low-resource language by Arabic script in China, there are still many serious challenges in these research areas by natural language processing. This paper standardized the processing coding and storage scheme of Kazakh corpus, then constructed Kazakh Language Corpus (KzLC), which lay the foundation for further research on syntactic analysis etc. of Kazakh language processing. Aiming at frequency issue of Kazakh language, this paper focused on relation of Zipf's law of power law in Kazakh word, which is based on frequency statistic of the word. On the basis of frequency statistics of Kazakh words from Kazakh textbooks, this research came up worth word information analysis and statistic method based on corpus, which revealed language rule and phenomenon among Kazakh words information.

Keywords: Kazakh language, statistics, corpus linguistics, word frequency, Information retrieval, morphological analysis.

Introduction

Natural language processing has become one of the significant technologies during the information technology development among different countries and nations. Frequency statistic word are important tasks in Kazakh natural language processing research, A corpus-based approach to natural language processing research has become very popular, A corpus-based to Morphological analyzers have been developed for different languages.

Kazakh Language belongs to the Turkish Language group in the Altaic language family, and it is an agglutinative language with word structures formed by adding derivational or inflectional affixes to root words. There are three different scripts of Kazakh characters all around the world: Cyrillic letter is used in Kazakhstan; P. R. China use Arabic letter, and Latin letter is widely used other country.

Previous work done on Kazakh language processing research such as Altenbek has designed the Kazakh corpus for Kazakh Arabic letter [1]. Makhambetov have designed the Kazakh Cyrillic letter corpus compilation process (Makhambetov et al. 2013). Washington et al. have established Finite-state morphological transducers for Kazakh Cyrillic letter [2]. Doszhan introduced about creation of all Turkic national corpus problem [3]. According to our knowledge, there is no research on frequency statistic by Morphological feature for Arabic letter Kazakh. This paper is a first time to do research work for frequency statistic of Arabic Kazakh Word based on corpus by Morphological feature.

In this paper, the research mainly focuses on a corpus-based Frequency Statistic of Kazakh Word at Morphological feature, which are the most difficult aspect of Kazakh natural language processing using the statistical method by Arabic script in P. R. China. Our research project has shown the corpus-based approach to processing of word frequencies, then Kazakh corpus is design consideration and building the corpus. Also the result expresses the relation of frequency of the Kazak word, which is based on the corpus from mainstream newspaper media and Kazakh school Textbooks. and the resulting Kazakh word frequency distribution accords with law of Zapf. Kazakh language is different from English and other languages that have been studied in corpus, as frequency statistic of words is an important part of Kazakh language processing. This research results not only created corpus resource for further information processing of Kazakh language, but also provided corpus data to Kazakh language linguistics research. This exploration is the basis task of machine translation, speech recognition, information retrieval and many other application developments in the Kazakh language.

Related work

Since the first corpus of Brown University was established in 1963 to 1964, corpus linguistics has become a important tasks in natural language processing field. The purpose of American Brown corpus is to study the modern American English, using the principle of system to collect 1 million words English text,

based on the use of rules and 86 kinds of grammatical markers of automatic POS tagging. The modern British English LOB corpus in 70s, also collected 1 million times, using 133 kinds of grammatical markers, it is POS using CLAWS (Constituent Likelihood Automatic Word-tagging System Constitute) achieve automatic part of speech tagging system by statistical information. In the past year, many researchers have been constructed their own language corpus, such as Korean National Corpus, Turkish National Corpus [4], Russian National Corpus, Chinese Peking University Corpus.

Study on morphology analysis of based-corpus can not only large-scale real language, and language specific qualitative explanation, corpus analysis provides a new research platform based on the use of language, language can be analyzed from the characteristics of the phenomenon of word frequency and syntactic language. The dictionary is written based on the corpus, collocation can search for specific words; vocabulary corpus based development, can survey the vocabulary of grammatical features, using feature; corpus based techniques can provide language learners with examples of language analysis.

Corpus Design Considerations

Corpus is a collection of some language texts sored in an electronic database. And it is the basic resource of natural language processing with statistical language model. The corpus plays an important role in knowledge acquisition, and corpus construction is a sign of corpus size and corpus selection.

The normative principles of constructing the Kazak Language Corpus.

Since 2009, we have constructed a word level corpus of Kazakh language, which has been continuously improved, modified and expanded in the past years. It is particularly important to find out the standards for the initial construction of the corpus, which will affect many of the later research results. Therefore, we carry out the following research.

(i) The objectives and collection criteria for source material: Quality is the lifeline of the corpus, emphasizes the universality and normative analysis; corpus obtained can basically summarize the whole or part of the Kazakh language specified characteristics as a representative corpus. So, it is to identify the principal aspects of corpus creation and the main decisions to be made.

(ii) Corpus size: The size of the corpus in relation to statistical data are reliable, and scarce resources belonging to the Kazakh language, source of corpus is less, scale is larger, the construction difficulty and cost more. Kazakh language is a low-resources, we select the corpus from mainstream newspaper media and Kazakh school Textbooks.

(iii) The processing and depth principles: Processing materials include text format processing and text description. According to the Kazakh characteristics, to provide users with data processing depth information for the purpose of Kazakhstan linguistics, corpus processing specification includes: the letter frequency statistics, word frequency statistics, word segmentation, word formation of additional components, the stem or POS tagging etc..

(iv) The encoding, description, storage format: According to the Kazakh language corpus source, input method and format inconsistencies, to facilitate unified management and sharing of resources, follow by the XML language and UNICODE encoding form for the corpus storage format, the Kazakh corpus description information to achieve the complete word tagging specification, is beneficial to the understanding and application of language resources and sharing. Tagged corpus is stored in XML file and TXT form separately.

Text documents description. In the initial version of our corpus, we used the data start from January 1, 2008 Xinjiang Daily (Kazakh version) for the Kazakh Language Corpus (KzLC) to whole year. The corpus consists of raw texts and POS tagged XML format texts. XML annotation content and text structure information as follow:

Title: sub- title

<TITLE><Subtitle></ Subtitle> </TITLE>

Paragraph

Sentence <s></s>

Word <W></W>

The word in the corpus annotation specification: adding the properties for a part of speech-POS, stem, affix, unknown word etc. In this experiment, a corpus based storage format: UNICODE as a corpus character encoding, using XML language as storage format; pure text or database form.

The Kazakh word corpus POS annotation: The Kazakh word tagging corpus is labeled various word level information based on the corpus of the language materials, in order to solve POS tagging of Kazakh lexical analysis, this study proposed a Kazakh word tagging Standards including POS tagging, stem tagging,

affix tagging, multi-class word POS tagging as an attribute. Through these annotation information, more in-depth analysis of the language material, can obtain more knowledge about the Kazakh language, lay the foundation for information processing but also for the construction of Kazakh language.

Part-Of-Speech (POS) tagging is the process of assigning a part-of-speech label to each of a sequence of words. There are many different tag-sets for the parts of speech of a language Include n, num, int, v, adj, adv, pron, part, ono and q. Table 1 presents our POS tag sets.

Table 1

Kazakh POS tag sets

No.	Tag	Name in Kz	Description	No.	Tag	Name in Kz	Description
1	n	زات ەسىم	noun	8	part	شىلاۋ	Auxiliary word
2	num	سان ەسىم	numeral	9	ono	ەلىكتەۋىش	onomatopoeia
3	int	وداعاي	interjection	10	q	مولشەرلىك	quantifier
4	v	ەتىستىك	verb	11	F	كىرمەسوزدەر	loanword
5	adj	سىن ەسىم	adjective	12	abbrev	قىسقارتپ	abbreviate
6	pron	ەسىمدىك	pronoun	13	punc	تىنىس بەلگىسى	punctuation
7	adv	ۇستەۋ	adverb	14	sym	بەلگىسى	symbol

Recent years, according to the phrase recognition of Kazakh shallow syntactic parsing problem, we investigated the basic structure system of Kazakh basic phrase , and study determined the noun, verb and adjective phrase structure as following. The basic phrase symbols using the IOB annotation approach for Kazakh language basic phrase (Table 2).

Table 2

Kazakh phrase tag sets and examples

Phrase	Tag	Keyword	E.g.	Begin-tag	In-tag	Out-tag
Noun phrase	NP	Noun	التىن كۇز	B-NP	I-NP	o
Verb phrase	VP	Verb	مۇراتقا جەتۋ	B-VP	I-VP	o
Adjective phrase	ADJP	Adjective	تاپ- تازا	B-ADJP	I-ADJP	o
Pronoun phrase	PRONP	Pronoun	ونىڭ مۇراتى	B-PRONP	I-PRONP	o
Numeral phrase	NUMP	Numeral	سەگىز توعىز مىڭ	B-NUMP	I-NUMP	o
Adverb phrase	ADVP	Adverb	ەڭ تەز	B-ADVP	I-ADVP	o

Constructing Kazakh corpus: Before collecting the Kazakh Arabic texts, a plan was made to decide on the start size of our corpus and text types. So, In order to achieve the balance and typical of the distribution of Kazakh language corpus, there are two main corpora for our research, one is use the 'Xinjiang daily' Kazakh version electronic text data whole 2008 year. The corpus includes news, economy, science and education, sports, Kazakhstan customs five columns. Another is the Kazakh Arabic textbooks from primary school, middle school to senior high school in Xinjiang, China, it is a classic and standard Kazakh teaching textbook. Through the construction of a few years, these two corpora and some other materials constitute our Kazakh language corpus – Kazakh language corpus – KzLC. Many text collection was done manually because Kazakh language is low resource language, and our task was to fine Arabic Kazakh text resources and obtain their copyright permission, then convert the different input text to our standard Unicode text format. For example, the raw corpus is media newspaper webpage format, and needs to convert into plain text corpus according to the 'year/month/date', and should remove junk information in webpage format, only retaining the effective text information content, and the converted file format for the TXT file.

Based on the Corpus evaluation principles of the standard, typical, structural and balanced, our work adopts the method of human-computer interaction and statistics to construct the annotated corpus. This construction tasks include dictionary resources, words and phrases annotation by syntactic level in Kazakh language corpus. The analysis of word statistics information based on corpus

Study on word statistics information analysis based on corpus includes many aspects, the natural language information statistical corpus based analysis of lexical distribution analysis on keywords based analysis of various circumstances analysis, the analysis of corpus based vocabulary research, including frequency collocation research, dictionary compilation, new words and popular words etc.. This paper focuses on the word frequency, the Kazakh word length and different stage, the Kazakh word frequency is consistent with the linguistic rules of Zipf's law.

In order to make a corpus based statistical analysis of Kazakh word information, the following terms are introduced. Frequency analysis refers to the distribution of data by means of frequency distribution tables and charts. Frequency refers to the ratio of the frequency of the word to that of the current corpus. The formula is as follows:

$$\mathrm{F}_i = n_i / N \times 100\,\%.$$

Here n_i — is the number of occurrences of the word; N — is the total number of occurrences of the corpus; F — is frequency; F_i — is the frequency of the word i.

Data of Corpus. In the experiment, there are two corpora for our research, First experimental data is «Xinjiang daily» Kazakh version of 2008 year electronic text data, Second experimental data is for analysis of word frequency use the Kazakh Textbooks. The Kazakh word frequency comply. Kazakh language has a complex morphological structure, similar to typical agglutinative languages. Words can be formed by long concatenations of morphemes with some order or semantic features. Kazakh words letters analysis with statistics: Firstly, We use the first experimental data for the following Initials letter statistical analysis of Kazakh words.

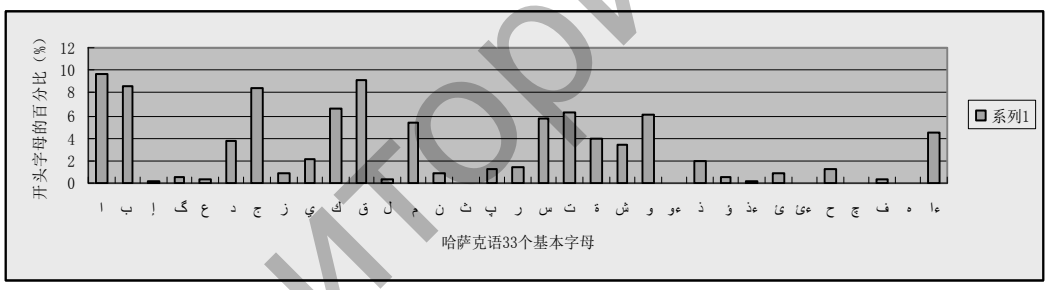

Figure 1. Initials letter statistics of Kazakh words

Figure 1 show the letter «ا» accounts for 9.71% of our total vocabulary, then accounted the letter «ق» and «ب» for 9.09 % and 8.66 %, the letter «ھ» is minimum number of words Accounts for only 0.019% of the total vocabulary only in interjection.

Then, we also use the first experimental data for the following all letter statistical analysis of Kazakh words. According to Figure 2, these three letters, “ا”、“ى”、“ن” appear lots of times in the textbook. However, three other letters, ۇ、ء、ھ did not appear to many times, which show the frequency of utilization of Kazakh letters.

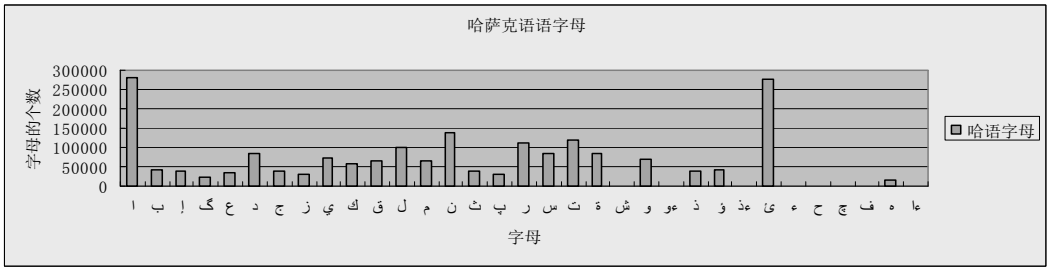

Figure 2. All letter statistics of Kazakh words

Finally, in order to do further explanation, for a comparative analysis of the statistics of the Arabic Kazakh alphabet statistical data and the Cyrillic Kazakh alphabet frequency statistics, the Kazakhstan scholar Makhambetov (Makhambetov et al., 2013). The Cyrillic Kazakh alphabet frequency statistics table shown in Figure 3. The Table shows that we are ranked first а, е, ы, н, «а, е, ы, н» four letters in Kazakhstan's Kazakh language is also ranked the first four. And the ranking of the «ә, һ» letters in Kazakhstan Kazakh language is also ranked low. This experiments show that the frequency of pure Kazakh voice letters all over the world are basically consistent.

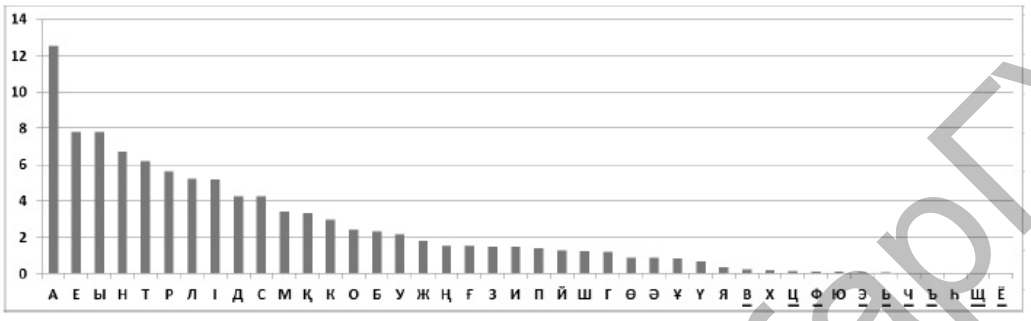

Figure 3. Kazakh Cyrillic letter statistics (Kazakhstan)

The Kazakh word frequency comply with Zipf's law of power law: The Zipf's law is named after the American linguist George Kingsley Zipf, it states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency Table 3. The Zipf's law formula is as follows:

$$f_r \times r = c.$$

Or

$$p_r = cr^{-\gamma},$$

here f — is frequency; r — is rank; c — is constant; γ — is parameter.

We use the first experimental data for the relationship between the frequency and length of words statistical analysis of Kazakh words. Length of the word is calculated based on numbers of letters in the word.

Table 3

The relationship between the frequency and length of words

No.	Length frequency	№	Length frequency	№	Length frequency	№	Length frequency
1	2.0054	5	13.4878	9	8.9305	13	1.6849
2	2.8539	6	13.8508	10	5.7958	23	0.0355
3	7.5870	7	13.1275	11	5.0928	26	0.0257
4	8.8262	8	10.9275	12	2.8579	30	0.0146

As shown in Table 3 proportion of three to nine letters words among all Kazakh words is 76.7373 %. The major part is five to seven letters words, which is 40.4661 %. Figure 4 is the relationship between length and frequency. It indicates three to nine letters words is the majority. And they are followed by ten to eleven letters words, which is 10.8886 %. One to two and twelve letter words have a proportion of 7.7172 %. Last part is thirteen to thirty letters words, which is 4.9989 %.

This experiment shows relationship between length and frequency, which prove that frequency of Kazakh words follows Zipf's power law. This result also indicates three to nine letters words are most words in Kazakh version of Xinjiang daily news. There are also some long words, which matches readers' reading habit and vocabulary level.

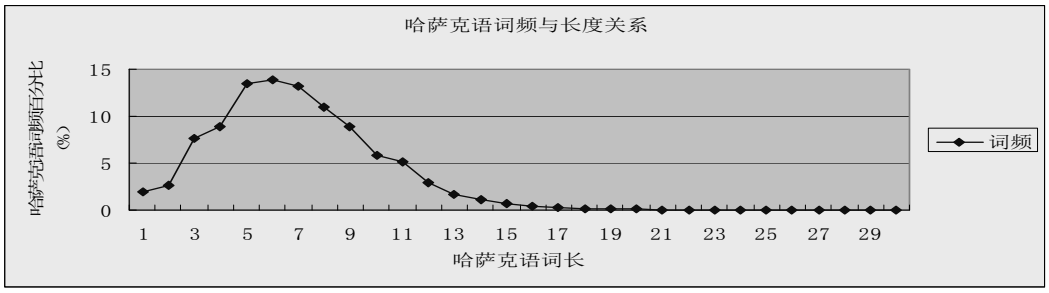

Figure 4.The relationship between the length and word frequency

The relationship between the length and frequency data from textbooks corpus in Figure 4. As about that the relationship between length and frequency of all word has a remarkable characteristic is to the left, which means most of the Kazakh word length is short, E.g. On the one hand, The length from 5 to 8 accounting is 51.3936 % of all words, from 13 to 30 accounting is only 4.9289 % of all words, on the other hand, Dragging a long «tail» is the another big characteristic of power law.

Statistical analysis of Kazakh words, stems, suffixes: We use the second experimental data for the following Kazakh words, stems, suffixes statistical analysis of Kazakh words in primary school, junior high school and high school.In order to explore Statistical analysis of Kazakh words,here word,stem and suffix are a kind of word segmentation unit in our corpus for non recurring terms, which is, the number of Kazakh word, stem and suffix except the punctuation mark and the English language, as shown in Figure 5.

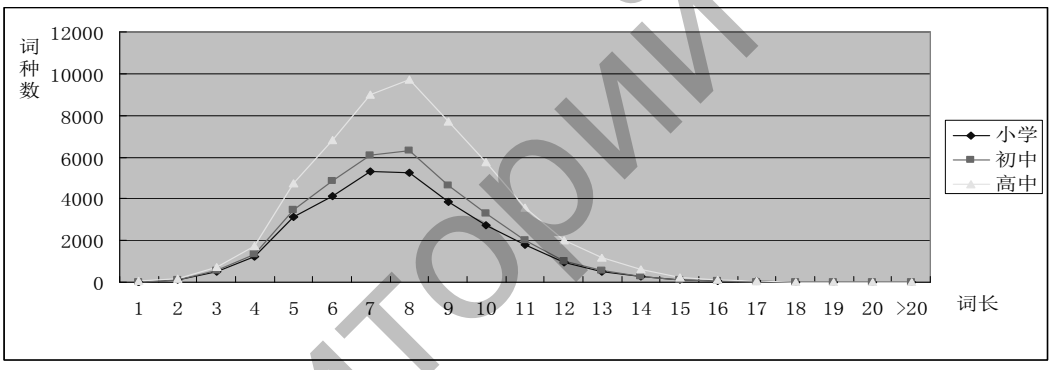

Figure 5. The relationship between the word of length and frequency in Kazakh textbooks

Figure 5 shows that words in three different taxtbooks are composed of 1 to 20 charaters.The Words composed of less than 3 or more than 15 charaters are used less commonly. In primary school, junior high school, and high school textbooks, word composed of less than 3 or more than 15 characters account for 2.07 % and 0.7 %, 1.91 % and 0.67 %, and 1.69 % and 0.9 % respectively.

In order to explore Statistical analysis of Kazakh stem for non recurring terms,we analyze the length of stem in primary school, junior high school and high school, as shown in Figure 6.

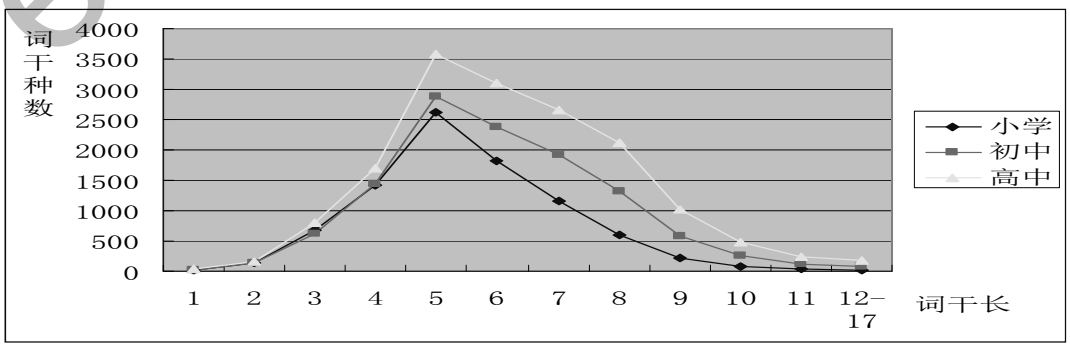

Figure 6. The relationship between the stem of length and frequency in Kazakh textbooks

Figure 6 shows that commonly used word stems in three different taxtbooks are composed of 4 to 8 charaters. 96.67 % (1505) of the word stems are composed of 4 to 8 charaters in primary school Kazakh language textbook. 94.84 % (11,167) of the word stems are composed of 3 to 9 charaters in junior high school Kazakh language textbook. 93.13 % (14,976) of the word stems are composed of 3 to 9 charaters in high school Kazakh language textbook. Word stems composed of less than 3 or more than 9 charaters are used less commonly.

This study summarizes the length of the word suffix used in three different textbooks for non recurring terms as shown in Figure 7.

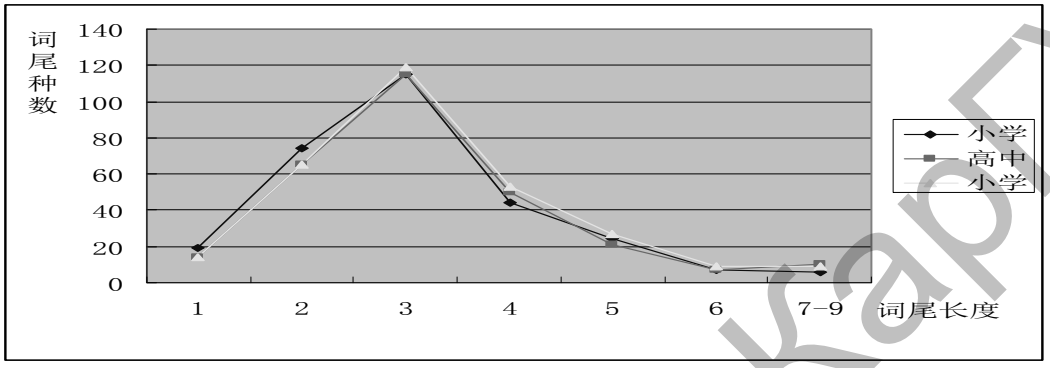

Figure 7. The relationship between the length and frequency of ending in Kazakh textbook

Figure 7 shows that the common suffix for words used in the three textbooks are usually composed of 1 to 5 characters. 95.49 %, 94.97% and 93.92 % of the suffix types appear in primary school, junior high school, and high school textbooks, respectively. There is little difference among the number of suffix types used in textbooks for different levels.

Figure 7 shows the similarity among the distribution of the endings in the words used in three different Kazakh textbooks, which indicates the characteristics of closed endings of Kazakh words, and the relatively stable word choice in Kazakh language. Finally, the analysis on the longest Kazakh words searched from the three Kazakh textbooks demonstrate that there tends to be little difference among the length of the longest words, as shown in Table 4.

Table 4

Instances of the longest Kazakh words

Resource	The longest word	Length
Primary school	سۋيسپەنشىلىكتەر بۇدى	20
	تازالانبايتىندىقتان	19
Junior high school	كورنەكتىلەنەتىندىگىنەن	22
	ورنالاستىرىلاتىندىقتان	22
High school	تۇجىرىمدالاتىندىقتان	20
	كورنەكتىلەندىرىلەدى	20
Commonly used word	قاناعاتتاندىرىلماعاندىرىلماعاندىق	23

Note. Length of the word is calculated based on numbers of letters in the word.

Table 4 presents the length of the longest words used in Kazakh textbooks for different levels. In general, there are 23 characters in the longest common Kazakh words. The analysis on the words used in the textbooks in Xinjiang, China shows that the longest words are composed of 19 to 22 characters. This stability of word choice in textbooks for different levels further proves the stability of Kazakh vocabulary.

Frequency of word is used as a measurement unit in language statistic. It indicates the number of times a word is presented in the corpus. It also includes the number of times the word is repeated. This section utilizes the following statistical analysis to the words in textbooks. These statistics reflect the breadth and depth of the contents of the textbook as shown in Table 5.

Table 5

The statistics of words usages in Kazakh textbook

Textbooks	Whole word	Word*	Suffix*	Stem*
Primary school	162 818	29 962	289	8 822
Junior high school	214 117	34 797	282	11 774
High school	336 475	54 202	296	16 808

Note. Here word*, stem*, suffix* are non recurring terms.

The frequency of words, the types of words and stems are increasing in these three textbooks, which indicates that the vocabulary of students' learning and accumulation is also increasing. The result shows that there is a big difference in numbers of words, however, small difference is types of word and stems, which indicates that the use of Kazakh dialect stem is relatively stable and word formation ability of the language is strong. And also it proved that in lexical system of Kazakh language the majority is gelling words with stem and suffix. Numbers of suffix in three textbooks are similar, which indicates closeness of Kazakh suffix. Figure 8 shows the relationship between them.

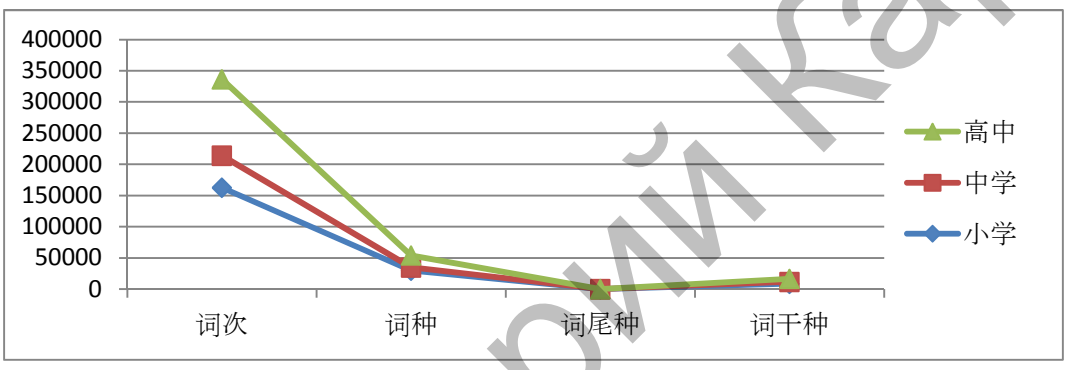

Figure 8. The statistics of words usage in Kazakh three textbooks

Statistics of high frequency words in each stage: In order to investigate the range of words used in primary schools, junior high schools and high schools, this paper enumerates the frequency of vocabulary in each stage, and lists the top 500 high frequency words for statistical analysis. Table 6 shows the first ten high frequency words Chart. From the analysis of Table 6, we can see that, regardless of the learning stage, the high frequency words of the Kazakh language are almost unchanged, and the highest frequency words are «بىر» (exactly one). And among the ten high frequencies words, eight of them are the same, just the frequency is different, which indicates that Kazakh language has a considerable stability.

Table 6

Part of most high frequencies words in each stage for textbooks

Primary school		Junior high school		High school	
word	frequency	word	frequency	word	frequency
بىر	1972	بىر	2143	بىر	3263
مەن	1334	دا	1949	دا	2783
دا	1303	دەپ	1724	مەن	2442
دەپ	1166	مەن	1626	دەپ	2329
ول	1000	دە	1494	دە	2182
دە	990	ول	1221	بۇل	1509
ەكەن	742	بۇل	917	ول	1473
بۇل	707	ەدى	913	بولادى	1197
وسى	603	ەكەن	842	دەگەن	1192
ونىڭ	576	وسى	824	وسى	1178

The following table is a high-frequency word statistics of Kazakh masterwork. Kazakhstan scholars in the literature did statistical analysis on the vocabulary of Kazakh masterwork Road of Abai. They list the top 500 high-frequency vocabulary, which the first shown in Table 7.

Table 7

High frequencies words in 《ABAY road》

Rank	Word	Frequency	Rank	Word	Frequency	Rank	Word	Frequency
1	**ده(v)**	9828	6	**دا**	5653	11	**وسى**	4379
2	**بول**	9341	7	**كەل**	5175	12	**ايت**	4301
3	**ه**	7844	8	**بۇل**	4696	13	**سول**	4175
4	**اياي**	6747	9	**ال**	4546	14	**ول**	4117
5	**ۇز**	5747	10	**ده**	4506	15	**بىر**	3896

Comparison of Tables 6 and 7 shows that there are five words with the same high frequency, just the frequency is different. The highest frequency word is «بىر» (meaning: one). Based on this, we further proved the stability and universality of the Kazakh vocabulary, and explained the unity of the Kazakh language.

Conclusion

In this paper, the research mainly focuses on a corpus-based Frequency Statistic of Kazakh Word at Morphological feature, which are the most difficult aspect of Kazakh natural language processing using the statistical method by Arabic script in P. R. China. This paper standardized the processing coding and storage scheme of our corpus, then constructed Kazakh Language Corpus (KzLC). The Kazakh word tagging corpus is labeled various word level information based on the corpus of the language materials, in order to solve POS tagging of Kazakh lexical analysis, this study proposed a Kazakh word tagging Standards including POS tagging, stem tagging, affix tagging, multi-class word POS tagging as an attribute.

Our research project has shown the corpus-based approach to processing of word frequencies, this study has been completed statistical content: Kazakh words starting letters with statistics, the Kazak of word frequency statistics, the Kazakh word length statistics, all the relationship between the length and frequency of the word. The experimental results illustrate the Kazak inner link between word frequency, and at the same time to verify the Kazakh word frequency comply with Zipf's law of power law.

In the experiment, we used the data of 2008 year of Xinjiang daily (Kazakh version) and Kazakh Textbooks from primary school and junior high school to high school in Xinjiang.by our Kazakh Language Corpus (KzLC).

For the future work we will plan to construct the syntactic annotation treebank of Kazakh language, which has been continuously improved, modified and expanded in the recent years. Then we have plan to construct a semantics annotation treebank of Kazakh language.

This research work is supported by Natural Science Foundation of P.R. China (NSFC), under grant No. 61363062, No.61063025 No.61572151 and other No. NMLR 201601.

References

1 Altenbek, G., Dawel, A., & Muheyat, N. (2009). A Study of Word Tagging Corpus for the Modern Kazakh Language // *Journal of Xinjiang University*, 26(4), 394–401.

2 Makhambetov, O., Makazhanov, A., Yessenbayev, Zh., Matkarimov, B., Sabyrgaliyev, I., & Sharafudinov, A. (2013). Assembling the Kazakh Language Corpus, In Proceedings of the Conference on Empirical Methods in Natural Language Processing 2013 (EMNLP 2013). Association for Computational Linguistics, 1022–1031.

3 Doszhan, G. (2013). Problems of Creation of the All-Turkic National Corpus. International Conference on Information // Business and Education Technology, 1018–1023.

4 Yesim, Aksan. & Mustafa, Aksan. (2009). Building a national corpus of Turkish: Design and implementation, Working papers in corpus-based linguistics and language education. Tokyo: Tokyo University of Foreign studies, 299–310.

Г. Алтынбек, Кс.Л. Ванг

Қазақша сөздерді қолдану жиілігінің статистикалық зерттеулері

Мақалада қазақ тіліндегі сөздердің қолданылу жиілігін статистикалық тұрғыдан зерттелді. Қазақ тілі агглютинативті тілдер тобына жататындықтан, табиғи тілді өңдеу саласында зерттелмеген мәселелер өте көп. Авторлар қазақ тілі корпусын кодтау және сақтау схемасын сипаттады. Сондай-ақ қазақ тілі корпусын (KzLC) жасауды ұсынды, бұл корпустық лингвистика саласындағы ғылыми зерттемелерге негіз болып, оның одан әрі дамуына ықпал етеді. Аталмыш зерттеу машиналық аударманың, сөзді (сөйлеуді) анықтаудың, ақпараттық іздеудің, қазақ тіліндегі басқа да қосымшаларды жасаудың базистік міндеттері болып табылады.

Кілт сөздер: қазақ тілі, статистика, корпустық лингвистика, сөздің қолданылу жиілігі, ақпараттық іздеу, морфологиялық талдау.

Г. Алтынбек, Кс.Л. Ванг

Статистические исследования частотности употребления казахских слов

Поскольку казахский язык относится к агглютинативной группе языков, существует множество проблем в области исследования обработки естественного языка. Авторами описана схема кодирования, хранения, обработки создания корпуса казахского языка и создания корпуса казахского языка (KzLC), что закладывает основу дальнейших научных разработок в области корпусной лингвистики. Это исследование является базисной задачей машинного перевода, распознавания речи, информационного поиска и многих других разработок приложений на казахском языке.

Ключевые слова: казахский язык, статистика, корпусная лингвистика, частотность слова, информационный поиск, морфологический анализ.