UDC 811.512.122

G.Altenbek[1], X.Wang[2]

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

Automatic Extraction of Kazakh Language

This paper focuses on the automatic extraction and frequency of the Kazakh Language. With automatic extraction of Kazakh Common-use words as the goal, we used a formula based on improved words filed general usage to calculate the lexical general usage of Kazakh Common-use words. This general usage was calculated on the basis of the filed usage of traditional words. This enabled an improved method which has greater ability to rank the position of Kazakh Common-use words and to implement the statistical system of Kazakh lexical general usage. Based on the Common-use word properties as Filed generality, Regional generality, and Time generality, we used statistical methods to investigate the general usage of Kazakh word. On the basis of frequency statistics of Kazakh words, we derived a formula for Kazakh lexical general usage. We extracted Common-use words in terms of the O_K value of lexical general usage. Experimental results show that the improved calculation formula has greater ability to rank word position than the traditional method used to extract Kazakh Common-use words.

Key words: kazakh word, lexical general usage, automatic extraction.

1 Introduction

Automatic extraction and word frequency are an important parts in natural language processing research. In a national language system, vocabulary is the basic carrier of language information and the most active, vital elements in the language system [1, 2]. Natural language processing research studies how to program the computer to understand the languages of human beings and develop relevant applicable systems. Common-use words are those used most frequently in everyday language and are a relative and stable open set.

Kazakh Language belongs to the Turkish Language group in the Altaic language family, it is written from right-to-left in the Arabic script in P.R.China. The Kazakh language have a rich vocabulary, and the research of Kazakh common-use word conforms to the social development. To make statistical analysis of a lot of different styles of written and spoken materials, and to use computer technology to study the Kazakh vocabulary in scientific research, it is important to identify the Kazakh common-use words [3, 4].

In this paper, we have used three features of common-use words that is, filed generality, time generality and regional generality. Using quantified 'lexical general usage' to indicate the extent of the general vocabulary, we applied the statistical method to investigate the general level quantity of Kazakh words in the mass media, then achieved the goal of automatic extraction system of Kazakh common-use word, which is based on the corpus from mainstream newspaper media.

2 Kazakh common-use word and its characteristics

2.1 Kazakh common-use word. Common-use words are those words which are used by people on a daily basis, not easy to change and relatively stable. Common-use words are the core of language. They have a close relationship to the daily life of the people and carry over between generations. They include terms for natural phenomenon, livestock, family name, parts of the human body, seasonal, activities number, work tools and so on. From ancient to modern society, the Kazakh people have been engaged in livestock husbandry, and have had a nomadic life, so the Kazakh common-use words also contained a large vocabulary about animal husbandry.

In the Kazakh language it is not very easy to divide common-use words out from the whole vocabulary. Even a few inaccuracies in the method used to determine common-use words can cause great difficulties. As a result, so far there is still not a scientific quantitative criterion for Kazakh common-use words.

Undeniably, common use words have historical stability, at the same time, they also slowly change. However, this does not preclude us from studying the features of common usage, time stability and word-formation ability for Kazakh, which is based on the vocabulary of Kazakh corpus, and investigating the use of these words in popular media. These inquires provide the basis for automatic extraction of Kazakh common-use words.

Therefore, refer to other language common-use words definition, the definition of Kazakh common-use words is the following: the use of Kazakh words in daily communication, mainly the words which have high frequency usage and high degree in each field, each local interval and each period.

2.2 Common-use word's characteristics. According to the definition of common-use word, we can get the common-use word's characteristic of generality, mainly lie in the following three aspects:

(1) Field generality: common-use word has the character of widespread use in all fields and trades;

(2) Regional generality: common-use word has the character of communicating wide spread with the people of different regions;

(3) Time generality: common-use word has the character of time stability.

Based on the three characteristics, we use the 'lexical general usage' indicators to quantize the description of these characteristics. Using the statistical method to investigate the general level of Kazakh universal vocabulary, which used by the mass media, in order to realize the goal of common-use word's automatic extraction for the Kazakh mainstream newspaper media based on the corpus.

3 Automatic extraction of Kazakh words

Kazakh common-use words have three characteristics: filed generality, time generality and regional generality, and use these characteristics corresponding to the field general usage, time general usage and regional general usage of quantitative indicators to measure the degree of general vocabulary.

3.1 Research on quantitative analysis method of field general usage. Field general usage of words is used to measure the general level in each field of circulation of language, which is the quantitative index of commonly used words degree. The computation formula not only should examine the frequency of a word, but also should consider whether words in different areas, different texts and the distribution of the fields are uniform. Mainly include two kinds of qualitative and quantitative investigation ways:

(1) Qualitative investigation: obtain all the fields of common words relying on fields' intersecting words.

(2) Quantitative investigation: calculate the word's field general usage according to frequency and distributed or other situation in each domain.

Qualitative investigation ways are relatively simple, for the field distribution of vocabulary is also clear at a glance, and its shortcoming is ignoring the words frequency—an important indicator to measure the common degree of words, so this kind of investigation ways only as the assistant to provide reference. We will use quantitative calculation ways to analyze the field general usage of words.

3.1.1 The traditional calculation method of Kazakh words' field usage. In order to measure the degree of commonly used words, the traditional method is used in addition to calculate the use frequency of words, even integrated considerate the calculation words with different areas of the text and area. The calculation procedure is:

(1) Calculate word frequency F_k:

F_k stands for total appearing frequency in the corpus of word *k*.

(2) Calculate the usage U_k of word *k*:

Use the improved calculating format of Chang Baoru's word usage:

$$S_k = \sqrt{\sum_{i=1}^{n} (N_k^i - N_k)^2 / n}.$$

$$D_k = 1 - S_k / (N_k \times (n-1)^{\frac{1}{2}}) \quad (0 \le D_k \le 1).$$

$$DI_k = \frac{P_k + L_k \times C_1 + C_2}{P + n \times C_1 + C_2} \text{ (Word comparative frequency} < 0.0001).$$

$$DE_k = \frac{1}{2} DI_k + \frac{1}{2} D_k \qquad \text{(Word comparative frequency} \ge 0.0001).$$

Calculating formula of words fields usage:

$$\text{Words usage } U_k = DE_k \text{ (or)} \times F_k \text{ (take integer)}.$$

Among them, D_k stands for the spread coefficient of word k. DE_k stands for the high frequency band of spread coefficient of word *k*. DI_k stands for the low frequency band of spread coefficient of word *k*. N_k^i stands for the relative frequency of word k in the field of I.N_k stands for the total relative frequency of word

k in all fields. P_k stands for distribution number of articles of word k. C_1 and C_2 are pending constant. *P* stands for the total articles of corpus. *N* stands for classification number field of corpus. Obviously, $1 \le P_k \le P$, $1 \le L_k \le n$. Make $n \times C_1 = P$, C_1 can be sure; Make $0.5 \le DI_k \le 1$, C_2 can be sure, C_1 and C_2 take a positive number.

Firstly need to calculate the constant value of C_1 and C_2. The value of C_1 and C_2 need to be proportional to the total number of articles, because the goal of this paper is to calculate the usage of each month, so *P* value is to process the total text of every month. By the validation of Xinjiang daily of Kazakh edition in 2008, this paper set s values as $C_1 = P/n$, $C_2 \approx 2 \times P$, make $C_2 = 1200$.

The above formula is considering the word frequency, word text spread and word field distribution and other integrated circumstance to obtain quantized index of commonly used features of words—-word usage, this is consistent with the goal of the words commonly used degree, so it can be used as the calculation formula of lexical general usage.

The traditional method has the following aspects of defects:

(1) Fixed word text distribution and field distribution together, it is needed to calculate two constants of value C_1 and C_2, and their values have a relationship to the total number of corpus and the domain classification number, so the formula is more suitable for application in a closed static corpus. Regarding the corpus which is used in this experiment, the total number of the corpus and domain classification numbers are always in dynamic status updates, so should not use the formula.

(2) Even in the total number of the corpus and the domain classification number under certain conditions, the value of C_1 can be uniquely identified, but the value of C_2 is required for a range of values, so the value of C_2 is not only, it also gives the calculation results bring uncertainty.

(3) Even though the field distribution characteristics in this formula have responses, their performance on the word general extent is relatively weak. But the lexical general usage requires that the field distribution parameters have more influence on measuring general degree of words.

3.1.2 The improved calculation method of Kazakh words' field general usage. In order to solve the weakness of traditional calculation formula of filed general usage, improve the calculateon method of filed general usage.

The calculation steps of improved filed general usage are as follows:

(1) Compute word frequency of field:

F_k stands for the total frequency of field classification corpus of word k.

(2) Calculate the usage of word *K* in text:

Use A.Juilland's formula to calculate the usage of words in text categorization:

$$S_k = \sqrt{\sum_{i=1}^{n} (N_k^i - N_k)^2 / n}.$$

$$D_k = 1 - S_k / (N_k \times (n-1)^{\frac{1}{2}}) \ (0 \le D_k \le 1.)$$

The usage of words $UL_k = D_k \times F_k$ (take integer values).

In Which, N_k^i stands for the relative frequency of word k in the field I.N_k stands for the total relative frequency of word k in all fields. N stands for the total text number of corpus. D_k stands for the coefficient of dispersion of word k. F_k stands for the frequency of word k.

(3) Calculate the field general usage of word U_k:

Use the Distributional Consistency (DC) to calculate the even degree of word in all areas. Calculation formula is:

The Distributional Consistency $DC_k = SMR/Mean$ $(0 \le DC_k \le 1)$ he definitions of SMR and Mean are as follows:

$$SMR = (\sum_{i=1}^{n} \sqrt{FK_i} / n)^2 \quad Mean = (\sum_{i=1}^{n} Fk_i / n).$$

The field general usage of words k $U_k = DC_k \times UL_k$.

In above formula, n stands for the number of field classification, and requires that the number of field of corpus is equal; FK_i stands for the frequency of word k in the field I. UL_k stands for the text usage of words k. DC_k stands for the even degree of word k in the field classification.

3.1.3 The analysis of experimental results of improved field general usage calculation method. As mentioned previously, calculation formula of field general usage can use the traditional calculation formula of word usage, but we discussed this formula's defects, especially the impact of the words' field distribution on field level of general quantitative calculation is not obvious, so this study proposed an improved formula.

In order to validate improved calculation formula, we select the Xinjiang daily of Kazakh edition in January 2008 as the corpus of test data, separately calculate the field general usage of traditional formula and improved formula, and make comparing analysis of this test.

Compared experimental analysis of frequency:

This paper's frequency is the unified tag of FKID and UKID, namely the corresponding in the words table according to sort of position after order of 'the frequency of words' and 'field general usage'.

UKID: FKID > 1, it is shown that the position of the words compare to the original FKID position is on the backward adjustment, the higher the value, the greater the adjustment of the word backward, then the text spread and the stability of the field distribution usage for the word is not good.

UKID: FKID < 1, it is shown that the position of the words compare to the original FKID position is on the forward adjustment, the higher the value, the greater the adjustment of the word forward, then the text spread and the stability of the field distribution usage for the word is good.

UKID: FKID = 1, it is shown that the position of this words is same as the original position of FKID, there is no effect for the range position of this word.

(1) The calculation formula of traditional field usage

Table 1 and 2 are the processing experimental results of the traditional calculation formula of field usage, cutting part of backward words and the adjustment forward Kazakh words.

Table 1

The backward words of field usage

Kazakh words	FK	Text NO	UK	FKID	UKID	UKID: FKID
بولدى	494	136	527.895608	276	630	2.282609
باستاماشى	246	50	129.506097	138	268	1.942029
قوعام	391	100	288.097844	231	477	2.064935
وسى	1986	361	827.966280	390	668	1.712821
حالىق	625	174	800.304715	311	659	2.118971
كوركەمونەر	240	50	119.456207	134	214	1.597015
جوق	525	149	518.871667	285	616	2.161404

Table 2

The forward words of the field usage

Kazakh words	FK	TextNO	UK	FKID	UKID	UKID: FKID
شارۋالار	368	62	109.049662	221	173	0.782805
باعاسى	232	25	34.756044	129	62	0.480620
بيگە	199	25	34.778509	104	64	0.615385
قىستاقتاردىڭ	335	62	110.70047	202	176	0.871287
ورىندادى	99	37	34.293415	28	24	0.857143
بارىنشا	340	87	113.13839	205	184	0.897561
تەگشە	99	25	34.176450	28	16	0.571429

The maximum value of UKID: FKID is 2.282609, the minimum value is 0.480626, so adjustment range of words' position for traditional computational formula is not big. Thus the text spread and field distribution does not bring about too much impact to sort result of words.

(2) The calculation formula of improved field general usage

Table 3

The backward words of the calculation formula

Kazakh words	Text number	FK	FKID	UK	UKID	UKID: FKID
دۇنيە	424	4957	1	859.3169442	5	5
شارۋا	374	3587	2	855.0248183	6	3
14	175	1905	9	137.4895981	322	35.77778
بەيلىق	265	1850	11	288.3072141	302	27.45455
مۇناي	162	1973	12	457.1887243	106	8.833333
25	112	1368	22	142.8154444	309	14.04545
ساۋدا	87	1036	37	105.8061292	889	24.02703
ولىمپيا	50	887	50	35.98305303	807	16.14

Table 4

The forward words of the calculation formula

Kazakh words	Text	FK	FKID	UK	UKID	UKID: FKID
بۇل	386	2932	4	871.8128425	1	0.25
اۇىل	49	113	985	300.4290849	142	0.144162
ەرى	100	336	215	533.8430249	39	0.181395
ەل	124	353	299	762.0626994	25	0.083612
مەدەنيەت	74	237	333	504.7022196	79	0.237237
جۇگگو	99	264	349	477.628399	94	0.269341
دەپ	87	256	361	520.531814	57	0.157895
ۇرىمجى	98	234	396	517.3909156	67	0.169192
شينجياڭ	74	232	405	493.3513744	86	0.212346
قازاق	74	288	425	517.3553599	68	0.16

Table 3 and 4 are the processing experimental results of the calculation formula of improved field general usage, cutting part of the backward words and the adjustment forward Kazakh words.

The maximum value of UKID: FKID is 35.77778, the minimum value is 0.083612, the adjustment range expands, so adjustment range of words' position for improved calculation formula is larger. Thus the impact strength of the text spread and field distribution of words is equal or not for the adjustment position of word and will increase.

The experimental results show that the traditional calculation formula does not change the adjustment position of word, therefore, the text spread and field distribution of words do not bring about too much impact on sort result of words. However, improved calculation formula changes the adjustment range of position of word a lot, thus the impact strength of the text spread and field distribution of words is equal or not for the adjustment position of word will increase, this is more accord with our recognition of the common-use words.

3.2 Calculation method of time general usage. The time general usage is the quantitative of words in the investigation time. It needs to observe whether the words in the inspection period are stable or not, namely the even degree of words in every month.

The procedure of time general usage is as follows:

(1) Calculate the monthly frequency of words:

F_k stands for total appearing frequency in every month for word k.

(2) Calculate the time general usage of word k:

Use Distributional Consistency (DC) to calculate equal degree of words in every month of the distribution in the investigation, the calculation formula is:

$$SMR = (\sum_{i=1}^{n} \sqrt{FK_i} / n)^2 \quad Mean = (\sum_{i=1}^{n} Fk_i / n).$$

The time general usage of word k T_k = SMR/Mean。 ($0 \le T_k \le 1$).

Above formulas, n stands for the number of months in investigation, the requirement of each month's corpus is equal; FK_i stands for the frequency in the month i.

3.3 Feature description of region general usage. Region general usage of word is the vocabulary's use situation in different regions of the media from the point of view of common time, that is, during the investigating time, the stability degree of vocabulary in different regions medial.

Region general usage is similar to the time general usage. They are also investigating the equality degree in different distribution of the classification system, which is the stable degree of using except that the difference is that time general usage is classified by months, but region general usage is classified by the different parts of the media.

Due to the present primitive data source of the experiment only having the Kazakh edition of Xinjiang daily, the regional representativeness is insufficient, so the influence of region general usage to lexical general usage temporarily is not taken into consider.

3.4 Calculation method of lexical general usage. As stated above, the description of the 'lexical general usage ' is considering the field general usage and time stability of words, and but does not take into account the influence of region general usage to lexical general usage.

Calculation formula of lexical general usage is:

Lexical general usage $O_k = T_k \times U_k$.

T_k stands for the time general usage of the word k. U_k stands for filed general usage of the word k. O_k stands for general level of word, the higher of the O_k value, the characteristic of common use and the stability feature of inspected time using performance are better.

4 Automatic extraction system

4.1 Selection and pretreatment of Corpus. In the experiment, use the 'Xinjiang daily' Kazakh version of 2008 year electronic text data to do statistics. In order to study the vocabulary field general, the original media corpus classification is needed. This paper will divide the raw corpus into 5 classes as political, economy, education, life and sports.

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.

4.2 Extract process of Kazakh common-use words. In order to calculate the lexical general usage of Kazakh words, firstly step the processed plain text format, and then extract the storage data, making frequency statistics for the extracted words. Field general usage and time general usage can be calculated when we get the word frequency. Finally we can calculate Kazakh lexical general usage. So we can extract the Kazakh common-use words according to the extraction rules when we get the Kazakh lexical general usage. The extraction of Kazakh common-use words mainly divides into two modules: preprocessing module and extraction module.

4.3 System implementation. According to the lexical general usage calculation method of Kazakh vocabulary, use C# language for system development, and recognize the Kazakh lexical general usage statistical system. Calculation process of Kazakh common-use words and system interfaces is as follows:

(1) Calculation of fields, each Kazakh text 'word' of the word frequency, Figure 1 shows the system interface of Kazakh word frequency statistic.
(2) Calculate text usage and field general usage in all areas' words monthly. Figure 2 shows the system interface of Kazakh word field general usage statistic.

Figure 1. Kazakh word frequency statistic system

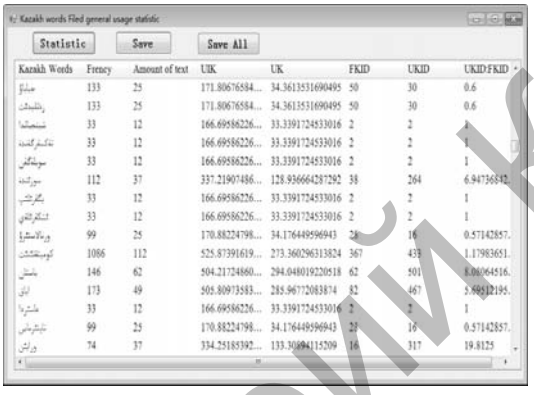

Figure 2. Kazakh word field general usage statistic system

(3) Calculation of annual Kazakh word time general usage and lexical general usage. Figure 3 shows the system interface of Kazakh word lexical general usage statistic.

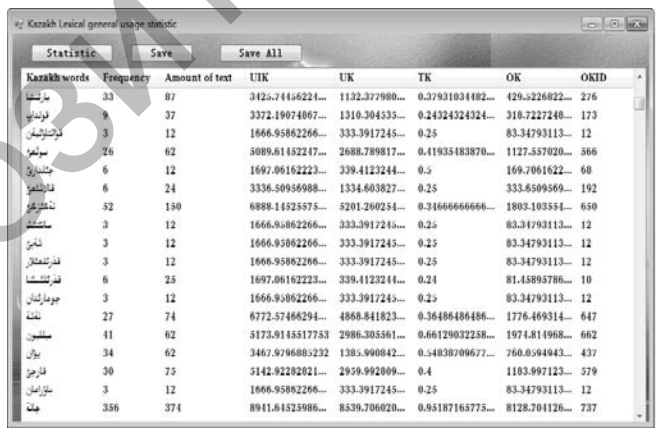

Figure 3. Kazakh word time general usage and lexical general usage statistic

4.4 Extract Kazakh common-use words. As mentioned before, lexical general usage shows that word's general level in mass circulation medium combines these statistical features: the word frequency, text distribution, field distribution, the use of time stability, and is the quantitative assessment index of these comprehensive performances. According to the 2008 annual 'Xinjiang daily' Kazakh version of the selected corpus words' lexical general usage O_K values from big to small order, cover the total corpus and identify 85 %-90 % words as common-use words.

5 Experiment results

According to the lexical general usage calculation method of Kazakh vocabulary, use C# language for system development, realize the Kazakh lexical general usage statistical system. In accordance with the Kazakh common-use words extraction rules, successfully extract Kazakh common-use words.

Table 5

The part of experimental results of the Kazakh lexical general usage

Kazakh words	FK	Text number	ULK	UK	TK	OK	OKID
جانە	356	374	8941.6452	8539.7060	0.95187	8128.7041	737
ءبىر	291	386	8892.6146	8711.3119	0.75388	6567.3362	736
كەرەك	140	187	8700.0821	7210.1816	0.74866	5397.9969	734
بۇكىل	68	125	8582.6442	7483.6931	0.54400	4071.1290	730
ۇلى	83	100	6933.5912	4877.2432	0.8300	4048.1118	729
ءار	140	275	8708.4042	7800.9534	0.50909	3971.3944	728
دەۋىر	97	173	8552.3111	6051.5284	0.56069	3393.0535	722
ءدۇنيە	120	261	8707.4002	8650.4981	0.39080	3380.6544	721
بولۋ	109	186	6980.2351	5176.0691	0.58602	3033.2878	716

From the experimental results, the extraction method is basically satisfactory, but the accuracy of lexical general usage still remains at rising further. The main factors affecting the accuracy are:

(1) In this experiment, the original corpus is divided into 5 categories: political, economy, education, life, sports. These 5 classes cannot be encountered by the newspaper media of various types of text.

(2) The current lexical general usage without considering the regional general usage on the general effect of words.

(3) At present, the corpus is only ‘Xinjiang daily’ Kazakh version of the 2008 annual year, and the size is greatly restricted.

6 Conclusion

This article used the definition of the Kazakh common-use words, and identified the Kazakh common-use words’ three characteristics: filed generality, time generality and regional generality, and used these characteristics corresponding to the field general usage, time general usage and regional general usage of quantitative indicators to measure the degree of general vocabulary. Based on the improved words filed general usage to achieve a Kazakh common-use words extraction, successfully extract Kazakh general vocabulary, and has good extraction effect, make up for the traditional words filed usage is insufficient, make words’ field distribution characteristics have more influence of the words’ commonly used characteristics.

In the article ‘lexical general usage O_k’ is considered the word field general usage and time general usage and put forward, not considering the regional general usage on the general effect of words.

In order to further improve the accuracy of Kazakh lexical general usage and increase the statistical data of persuasion, the next step work should be expanded corpus size, increased number of text classification, not limited to the current 5 areas, and should also consider the larger geographical scale circulation corpus, such as increased Kazakhstan's media corpus, regional general usage into ‘lexical general usage’.

This work is funded by the Natural Science Foundation of P.R.China (NSFC)(No.61363062, No. 61063025 and No.61272383), Science and Technology Research and Development Funds of Shen-zhen City (No. JC201005260118A).

References

1 *Wang Q., Park I., Zhang P.* Automatic Extraction of the Unlisted Terms In The Field of Information Technology Based on The Dynamic Circulation Corpus // Proceedings of IEEE. — 2003. — P. 452–458.

2 *Xiao-bing Zh.* A study on Recognition and extraction method of Contemporary Chinese Basic Vocabulary Based on Dynamic Circuit Corpus. Beijing: Beijing Language and Culture University, 2007.

3 *Hapar M., Altenbek G.* Design and Implementation of Kazakh Text Categorization System // Computer Engineering. — 2011. — No. 37(5). — P. 196–198.

4 *Abilhaye D., Altenbek G.* Study and implementation of Kazakh lexical scanner // Computer Engineering and Application. — 2008. — No. 44(19). — P. 146–149.

Г.Алтынбек, К.Ван

Автоматты түрде қазақ тілінен сөз, үзінді алу

Мақала қазақ тілінде автоматты түрде жалпылама лексикадағы қолданылатын үзінді алуға арналған. Дәстүрлі сөздерді қолдану негізінде жиі пайдалатыналатын сөздер алынды. Сөз қолдану жиілігі статистикасының негізіне сүйене отырып, біз қазақтілді лексикада қолданылатын формула жасадық. Сондай-ақ күнделікті қолданыстағы лексикалық мәні құнды деп саналатын сөздерді бөліп алдық. Тәжірибенің нәтижесі көрсеткендей, есепті формула сөздердің орнын бағалау, бұрыннан қолданылып келе жатқан дәстүрлі әдіске қарағанда, қазақ тіліндегі күнделікті қолданылатын үзінді алынатын сөздердің зор мүмкіншілігі бар екендігін көрсетті.

Г.Алтынбек, К.Ван

Автоматическое извлечение слов, отрывков в казахском языке

В статье представлен метод автоматического извлечения слов из общеупотребимой лексики казахского языка. Частота употребления слов вычислялась по количеству использования традиционных слов. На основе статистики частотности казахских слов была получена формула для использования казахскоязычной лексики. Извлечение слов проводилось исходя из ценностного лексического значения. Результаты эксперимента показывают, что по сравнению традиционным методом представленная формула вычисления дает больше возможности оценить частотный статус слов.