






# Experimental Verification of Collocation Detection Methods

Galiya S. Ybytayeva<sup>1</sup><sup>a</sup>, Nina F. Khairova<sup>2</sup><sup>b</sup>, Orken Zh. Mamyrbayev<sup>3</sup><sup>c</sup>,  
Kuralay Zh. Mukhsina<sup>3</sup><sup>d</sup> and Bagashar Zh. Zhumazhanov<sup>3</sup><sup>e</sup>

<sup>1</sup>Satbayev University, 22 Satpayev Street, Almaty, 050000, Kazakhstan

<sup>2</sup>National Technical University “Kharkiv Polytechnic Institute”, 2 Kirpichov Str., Kharkiv, 61000, Ukraine

<sup>3</sup>Institute of Information and Computational Technologies, 28 Shevchenko Str., Almaty, 050010, Kazakhstan

**Keywords:** Collocation, Corpus, Corpus Linguistics, Corpora, Association Measures.

**Abstract:** The article describes the results of a study to determine the correct phrases in the Kazakh language. The experiment consisted in the search and analysis of bigrams with frequent verbs, adjectives and nouns of the Kazakh language. Applying a statistical method to corpus material allows researchers to quantify the data obtained. The article provides an overview of MI, t-score indicators for calculating the strength of links within phrases, including their main characteristics. The purpose is to study the combinability characteristics of these lexical units, to correlate the results obtained on the basis of various association measures on different corpus, to compare the most popular association measures.

## 1 INTRODUCTION


In the process of penetration of modern information and communication technologies into all areas of science, in particular, into philological science, the popularity of using linguistic corpora of texts in the study of various aspects of the language is growing. In recent years, a whole range of methodological studies has appeared in the methodological literature on teaching schoolchildren and students the lexical and grammatical side of a foreign language using various linguistic corpora (Sysoyev, 2010; Chernyakova, 2012; Ryazanova, 2012). An analysis of this and other studies shows that the authors have reached a certain agreement on the conceptual content of the term “corpus linguistics”. It refers to an organized collection of texts selected and tagged according to a specific methodology and presented electronically.


The main attention in our research is paid to the corpus of parallel texts. In our study, we understand the corpus of parallel texts as a type of corpus linguistics consisting of a source text in one language and its translation into another language or languages.


This is a linguistic corpus of texts that allows you to study lexical connectivity or the phenomenon of word combinations in context.


Recently, in connection with the increasing need for automated systems, much attention is paid to the problem of automatic segmentation of word combinations in texts. There are various statistical indicators to evaluate the compatibility of words. Some dimensions are called associative measures or association measures. They allow you to calculate the strength of the connection between the elements of word combinations and are based on the frequency of these word combinations and the individual words included in them. Thus, it is possible to calculate some characteristics of the stability of lexical units, which allows them to be arranged on a conditional scale: from free combinations to phraseological units. In total, there are more than 80 measures to assess the strength of the connectedness of word combinations (Pecina, 2009).


The article is organized as follows. Section 2 is devoted to the literature review. Section 3 provides an overview of the statistical method, Section 4 presents the research methodology, and the final section discusses the research findings and suggests future plans.

<sup>a</sup> <https://orcid.org/0000-0002-4243-0928>

<sup>b</sup> <https://orcid.org/0000-0002-9826-0286>

<sup>c</sup> <https://orcid.org/0000-0001-7569-1721>

<sup>d</sup> <https://orcid.org/0000-0002-8627-1949>

<sup>e</sup> <https://orcid.org/0000-0002-5035-9076>

## 2 RELATED WORKS

Although the term “collocation” has recently come into regular use, it occupies one of the most important places in modern linguistics. In a broad sense, it is a combination of two or more words that tend to co-occur. Currently, collocations play a leading role in lexicographic practice (Atkins and Rundell, 2008; Kilgarriff, 2006). Recently, special collocation dictionaries are being created abroad and in Kazakhstan (Krishnamurthy, 2006; Smagulova, 2010; Zhanuzak et al., 2011).

However, existing dictionaries of regular expressions, firstly, do not contain their complete list, and secondly, they often do it in an insufficiently consistent manner. This is especially true for the Kazakh language. Therefore, the relevance of works on automatic detection of collocations from texts is undeniable.

Currently, we see several important application tasks that require automated methods for extracting collocations from large corpora of texts. In particular, these tasks include the creation of dictionaries and other lexicographic tools, the creation of ontologies, language learning, repair of linguistic processors, and information retrieval.

Let us briefly discuss the concept of the word combination. There are different definitions of this concept. In general, many definitions of collocation are based on the phenomenon of semantic and grammatical interdependence of phrase elements (Iordanskaya and Mel’chuk, 2007).

The term “collocation” in the Russian scientific literature was first used by Akhmanova (Akhmanova, 1996) in the Dictionary of Linguistic Terms. The first work in Russian linguistics devoted to the study of the concept of collocation in the material of the Russian language was the monograph of Borisova (Borisova, 1995).

Kozhakhmetova et al. (Kozhakhmetova et al., 1988) were worked on the problem of translation of correct word combinations from the Kazakh language to a foreign language without loss of meaning and national-cultural aspect. The scholars published a dictionary of some 2,300 regular expressions. It is effective to use in verbal translation, but we believe that it would be more effective if the regular word combinations were divided into meaning categories.

## 3 STATISTICAL METHOD

Nowadays the term “collocation” is widely used in corpus linguistics, in which the concept of collocation

is reinterpreted or simplified compared to traditional linguistics. This approach can be called statistical. Priority is given to the frequency of coincidences, so word combinations in corpus linguistics can be defined as statistically persistent phrases. In addition, a statistically persistent combination can be phraseological and arbitrary. In recent years, a lot of research and development on collocations has appeared, addressing both the theoretical aspects of a statistical approach to this notion and practical methods of phrase detection.

This is the emergence of a large representative corpus of texts, allowing to obtain reliable information on the frequency of a particular combination in the language as a whole. A high value of the frequency of matches seems to indicate the stability of the combination. However, this description is not sufficient to talk about the preferred combinability of certain words. Therefore, a number of statistical measures (called “association measures”) have been created to calculate the strength of the relationship between elements in a word combination. In general, these measures take into account both the frequency of matching and other parameters, primarily the frequency in a given corpus of each individual element.

However, statistics are not enough. The question needs to be answered as to what other requirements such statistically stable combinations should meet.

Most corpus managers are able to calculate the frequency of occurrence of words or word forms and the frequency of matches. Based on this data, there are many measures of association.

The total number of these dimensions is counted in dozens. The values of associative measures can be seen as indicators of the strength of the syntagmatic relationship between phrasal elements. See (Evert, 2004) for a description of the most common measures. MI, t-score is used more frequently than others. Some case managers allow the calculation of these measures.

The MI (mutual information) measure introduced in (Church and Hanks, 1990) compares context-dependent frequencies, such as randomly occurring words in a text, with independent frequencies:

$$MI(n, c) = \log_2 \frac{f(n, c) \cdot N}{f(n) \cdot f(c)}, \quad (1)$$

here:  $n$  – keyword (node);  $c$  – collocation;  $f(n, c)$  – frequency of occurrence of keyword  $n$  paired with collocation  $c$ ;  $f(n)$ ,  $f(c)$  – absolute (independent) frequency of keyword  $n$  and word  $c$  in the corpus (text);  $N$  – total number of word uses in the corpus (text).

If the value of  $MI(n, c)$  is greater than a certain value, then the expression can be considered statisti-

cally significant. If  $MI(n, c)$  is less than zero, then  $n$  and  $c$  are called complementary.

The t-score also takes into account the frequency of occurrence of a keyword and its combination, answering the question of how non-random the strength of the association between the word combinations is:

$$t - score = \frac{f(n, c) - \frac{f(n) \cdot f(c)}{N}}{\sqrt{f(n, c)}} \quad (2)$$

#### 4 IDENTIFICATION OF WORD COMBINATIONS BASED ON THE STATISTICAL METHOD

The aim of the work is a comparative analysis of different associative measures based on the corpus of the Kazakh language. In addition, the dependence of the results (the list of word combinations derived from the same measure) on the text material (text type) is investigated.

Our dataset includes a parallel Russian-Kazakh corpus, which has been developed over three years (Khairova et al., 2019), and an XML dictionary of synonyms with criminally related vocabulary (Khairova et al., 2021). The parallel Kazakh-Russian corpus includes texts from four news sites of the Kazakh information Internet space zakon.kz, caravan.kz, lenta.kz, nur.kz for the period from April 2018 to June 2021.

At the moment the volume of the parallel Kazakh-Russian corpus is 3000 texts in Russian and 3000 in Kazakh, including two thousand texts containing agreed Kazakh-Russian sentences.

We extracted the vocabulary for our XML dictionary of synonyms manually from the English, Ukrainian, Kazakh and Russian texts on criminal matters. Seven main thematic categories were identified for the terms, Movement, Traffic Accident, Injure, Offense, Arrest, Trial, PD. The choice of categories was due to the fact that the information resources from which the texts were taken contained the largest amount of data on the three criminal areas of “Police”, “Transfer”, “Crime” and their aforementioned subspecies. This made it possible to make our dictionary narrowly focused. All terms were also divided by parts of speech, i.e. only nouns, verbs, adjectives and word combinations were included in the dictionary. Figure 1 shows a fragment of the dictionary, which now includes about 650 basic words (over 320 nouns, over 100 adjectives, about 170 verbs and 40 word combinations) and over 2500 synonyms. It is currently still under active development.

Our study was based on the corpus of news texts “nur.kz”, “zakon.kz”, “patrul.kz”, “caravan.kz”, “inform.kz”, which includes 857 texts.

Table 1 contains data for 15 word combinations with word “police” sorted by value of MI parameter. The columns of the table, in addition to the word combination itself, show the following characteristics: Freq Word 1 & Word 2 – frequency of matching, Freq Word 1 – frequency of word combination, Freq Word 2 – key word, MI – MI value, T-score – t-score value.

The analysis of the data in table 1 (15 word combinations in total) shows that the ranks of the word combinations obtained using different indicators do not coincide.

It should also be noted that different dimensions affect the frequency of the words composing a word combination and the frequency of their combinability. Thus, MI is considered to be sensitive to low-frequency words, while t-score is useful for finding high-frequency word combinations.

We compared the automatically generated word combinations on different association indices with data from different dictionaries. The material served as collections of 2 nouns without homonyms (sozdik.kz) and 1 adjective (Kazakh-Russian, Russian-Kazakh Terminological Dictionary. Jurisprudence).

We call the above word combinations “correct”.

Below are graphs showing MI values, t-score measurements on the ordinate axis and bigram ranks on the abscissa axis. Black colour indicates “correct” word combinations from the dictionary “sozdik.kz” (4, 10 ranks) and “Kazakh-Russian, Russian-Kazakh terminological dictionary. Jurisprudence” (rank 7) indicates an additional phrase found in the dictionary.

The same tendency is observed for all obtained word combinations: the smaller the value of the measure, the greater the probability that these word combinations will not be registered as regular word combinations in dictionaries of the Kazakh language. Thus, we can say that the compatibility data given in dictionaries corresponds to the data obtained on the basis of associative measures.

As a result of the experiment, it seems important to identify phrases that are not registered in any of the dictionaries. The analysis of such word combinations shows that the bigrams located at the top of the list by degree of probability (sorted in descending order of one of the dimensions) turn out to be stable, so they can be included in the list.

As mentioned above, other statistical criterion methods based on linguistic models should also work. This idea has been adopted and implemented in the fa-

```

<term id="27">
<lemma lang="ru">сотрудник полиции</lemma>
<domain>PD</domain>
<synset lang="ru">агент полиции, сотрудник милиции, полицейский чиновник, офицер полиции, полиция
</synset>
<definition lang="ru">управление государственных служб и органов по охране общественного порядка
</definition>
<example lang="ru">Местные жители сообщали, что выживший сотрудник полиции родом из райцентра
Кыринского района</example>
<lemma lang="en">police officer</lemma>
<synset lang="en">policeman, patrolman, peace officer</synset>
<definition lang="en">a person whose job is to enforce laws, investigate crimes, and make arrests: a
member of the police</definition>
<example lang="en">But cadets and police officers deployed at the residence rapidly resorted to
sustained and excessive lethal force</example>
<lemma lang="ka">полиция қызметкері</lemma>
<synset lang="ka">полиция агенті, милиция қызметкері, полиция офицері, полиция, полицей</synset>
<definition lang="ka">қоғамдық тәртіпті сақтау, қылмыспен күресу, халық пен мемлекеттің
қауіпсіздігін қамтамасыз ету, қоғамдық және мемлекеттік құрылысты қорғау міндеттері жүктелген мекемелер мен
лауазымдардың және әскерлендірілген жасақтардың жиынтық атауы</definition>
<example lang="ka">Әлеуметтік желілерде полиция қызметкері өз тойында аспанға қарудан оқ атты деген
ақпарат тараған болатын</example>
<lemma lang="ua">співробітник поліції</lemma>
<synset lang="ua">агент поліції, співробітник міліції, поліцейський чиновник, офіцер поліції,
поліція</synset>
<definition lang="ru">орган державної влади, що займається охороною громадського порядку і боротьбою
з правопорушеннями</definition>
<example lang="ru">У Польщі співробітники поліції викрили три плантації марихуани на суму понад 2,4
мільйона злотих</example>
</term>

```

Figure 1: The fragment of the multilingual synonyms dictionary.

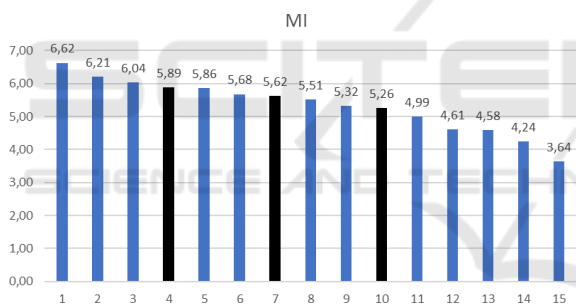


Figure 2: Values of the MI measure for collocations with the word "Police".

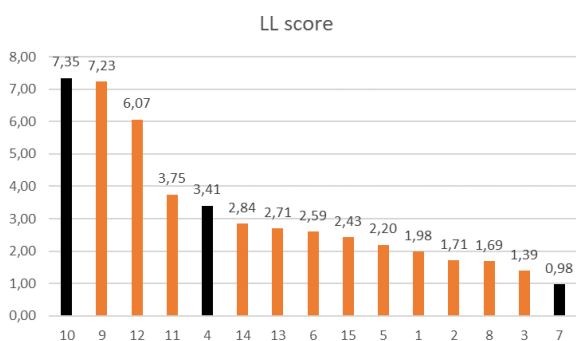


Figure 3: Values of the t-score measure for collocations with the word "Police".

mous Sketch Engine (Kilgarriff et al., 2004). It yields typical word combinations for a given keyword, on the one hand, due to a syntax restricting the compatibility of words in a given language, and on the other

hand, due to possible laws related to semantics and linguistic origin.

It turns out that there are few "correct" collocations, but this is because the vocabulary we have been relying on is too small, so it needs to be expanded. We can say that a new vocabulary is needed, which should contain various regular expressions.

The results of searching and identifying word combinations of this type are useful for lexicographers who know how to select different examples for dictionaries, and for linguists who study vocabulary and syntax in a certain aspect.

## 5 CONCLUSION

When comparing the phrases obtained using statistical methods with dictionaries, the same tendency is observed: the lower the value of the measure, the more these phrases are not recorded in dictionaries of the Kazakh language, and vice versa. Most of the phrases recorded in dictionaries are at the top of the list based on one of the measures of association. Thus, it can be said that the data on stable compatibility given in dictionaries coincide with the data obtained on the basis of measures of association, or, in other words, statistical measures of association better determine the real semantic-syntagmatic relations.

A comparative analysis of different association

Table 1: Values of associative measures for the word “Polisia”.

Nº	Collocation	Word 1	Word 2	Freq Word 1 & Word 2	Freq Word 1	Freq Word 2	Word in Corpus	MI	T- score
1	Patrúldik polisia	patrúldik	polisia	4	8	906	178645	6,62	1,98
2	Qarjy polisiasy	qarjy	polisiasy	3	8	906	178645	6,21	1,71
3	Polisia jasaǵy	polisia	jasaǵy	2	906	6	178645	6,04	1,39
4	Polisia bólimi	polisia	bólimi	12	906	40	178645	5,89	3,41
5	Polisiaǵa júginý	polisiaǵa	júginý	5	906	17	178645	5,86	2,20
6	Polisia shaqyrý	polisia	shaqyrý	7	906	27	178645	5,68	2,59
7	Áskerı polisia	áskerı	polisia	1	4	906	178645	5,62	0,98
8	Turǵylyqty polisia	turǵylyqty	polisia	3	13	906	178645	5,51	1,69
9	Polisia qyzmetkeri	polisia	qyzmetkeri	55	906	272	178645	5,32	7,23
10	Polisia basqarmasy	polisia	basqarmasy	57	906	294	178645	5,26	7,35
11	Polisia qyzmeti	polisia	qyzmeti	15	906	93	178645	4,99	3,75
12	Polisia departamenti	polisia	departamenti	40	906	323	178645	4,61	6,07
13	Polisiaǵa habarlasý	polisiaǵa	habarlasý	8	906	66	178645	4,58	2,71
14	Polisia basshysy	polisia	basshysy	9	906	94	178645	4,24	2,84
15	Polisia kóligi	polisia	kóligi	7	906	111	178645	3,64	2,43

measures carried out on a set of all data obtained for different word classes shows the following.

The MI measure can give the best average result. It makes it possible to distinguish between correct phraseological collocations as well as collocations in which proper names act as collocations, as well as low-frequency special terms. The disadvantages of using the t-score are primarily related to the fact that it determines the frequency with collocations, in particular with auxiliary words. Therefore, in order to “remove” the most frequent words for t-score, it is necessary to set up a list of stop words whose combinations are always at the top of the table: auxiliary words, pronouns or conjunctions. However, this also applies to other dimensions.

Whether statistical measures should be taken into account when searching for a lemma or a phrase remains an open question. The structural syntactic formulas and semantic constraints underlying the phrases also need to be taken into account.

In the future it is planned to test the effectiveness of the method on a large corpus.

## ACKNOWLEDGEMENTS

This work was carried out with the financial support of the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan (No. AR09259309).

## REFERENCES

- Akhmanova, O. S. (1996). *Slovar' lingvisticheskikh terminov*. Editorial URSS, Moscow.
- Atkins, B. T. S. and Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.
- Borisova, Y. G. (1995). *Kollokatsii. Chto eto takoye i kak ikh izuchat'*. Filologiya, Moscow, 2 edition.
- Chernyakova, T. A. (2012). *Metodika formirovaniya leksicheskikh navykov studentov na osnove lingvisticheskogo korpusa*. The thesis for the degree of candidate of pedagogical sciences.
- Church, K. W. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29. <https://aclanthology.org/J90-1003>.
- Evert, S. (2004). *The Statistics of Word Co-occurrences: Word Pairs and Collocations*. PhD thesis, Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, Stuttgart.
- Iordanskaya, L. N. and Mel'chuk, I. A. (2007). *Smysl i sochetayemost' v slovare*. Yazyki slavyanskikh kul'tur, Moscow.
- Khairova, N., Kolesnyk, A., Mamyrbayev, O., and Mukhsina, K. (2019). The Aligned Kazakh-Russian Parallel Corpus Focused on the Criminal Theme. In Lytvyn, V., Sharonova, N., Hamon, T., Cherednichenko, O., Grabar, N., Kowalska-Styczen, A., and Vysotska, V., editors, *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Systems (COLINS-2019). Volume I: Main Conference, Kharkiv, Ukraine, April 18-19, 2019*, volume 2362 of *CEUR Workshop Proceedings*, pages 116–125. CEUR-WS.org. <http://ceur-ws.org/Vol-2362/paper11.pdf>.

- Khairova, N., Kolesnyk, A., Mamyrbayev, O., Ybytayeva, G., and Lytvynenko, Y. (2021). Automatic multilingual ontology generation based on texts focused on criminal topic. In Sharonova, N., Lytvyn, V., Cherednichenko, O., Kupriianov, Y., Kanishcheva, O., Hamon, T., Grabar, N., Vysotska, V., Kowalska-Styczen, A., and Jonek-Kowalska, I., editors, *Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). Volume I: Main Conference, Lviv, Ukraine, April 22-23, 2021*, volume 2870 of *CEUR Workshop Proceedings*, pages 108–117. CEUR-WS.org. <http://ceur-ws.org/Vol-2870/paper11.pdf>.
- Kilgarriff, A. (2006). Collocationality (and how to measure it). In Corino, E., Marellò, C., and Onesti, C., editors, *Proceedings of the 12th EURALEX International Congress*, pages 997–1004, Torino, Italy. Edizioni dell’Orso. <https://euralex.org/publications/collocationality-and-how-to-measure-it/>.
- Kilgarriff, A., Rychlý, P., Smrz, P., and Tugwell, D. (2004). The Sketch Engine. In Williams, G. and Vessier, S., editors, *Proceedings of the 11th EURALEX International Congress*, pages 105–115, Lorient, France. Université de Bretagne-Sud, Faculté des lettres et des sciences humaines. <https://euralex.org/publications/the-sketch-engine/>.
- Kozhakhmetova, K. K., Zhaysakova, R. E., and Kozhakhmetova, S. O. (1988). *Kazakh-Russian phraseological dictionary*. Mektep, Almaty.
- Krishnamurthy, R. (2006). *Collocations*, pages 596–600. Elsevier, Netherlands, 2nd edition. <https://doi.org/10.1016/B0-08-044854-2/00414-4>.
- Pecina, P. (2009). *Lexical Association Measures: Collocation Extraction*. Studies in Computational and Theoretical Linguistics. Institute of Formal and Applied Linguistics, Prague. [https://ufal.mff.cuni.cz/books/preview/pecina\\_preview.pdf](https://ufal.mff.cuni.cz/books/preview/pecina_preview.pdf).
- Ryazanova, Y. A. (2012). *Metodika formirovaniya grammaticheskikh navykov rechi studentov na osnove lingvisticheskogo korpusa*. The thesis for the degree of candidate of pedagogical sciences.
- Smagulova, G. S. (2010). *Magynalas frazeologizmder sozdigi*. Yeltanym baspasy, Almaty.
- Sysoyev, P. V. (2010). Lingvisticheskiy korpus v metodike obucheniya inostrannym yazykam. *Yazyk i kul'tura*, 1(9):99–111.
- Zhanuzak, T., Omarbekov, S., and Zhunisbek, A. (2011). *Kazak adebietinin sozdigi. On bes tomdyk*. Almaty.