Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2017”

Moscow, May 31—June 3, 2017

AUTOMATIC COLLOCATION EXTRACTION: ASSOCIATION MEASURES EVALUATION AND INTEGRATION

Zakharov V. P. (v.zakharov@spbu.ru)

Saint-Petersburg State University, Saint-Petersburg, Russia

The paper deals with collocation extraction from corpus data. A collocation is meant as a special type of a set phrase. Many modern authors and most of corpus linguists understand collocations as statistically determined set phrases. The above approach is the basic point of this paper which is aimed at evaluation of various statistical methods of automatic collocation extraction. There are several ways to calculate the degree of coherence of parts of a collocation. A whole number of formulae have been created to integrate different factors that determine the association between the collocation components. Usually, such formulae are called association measures.

The experiments are described which objective was to study the method of collocation extraction based on the statistical association measures. We extracted collocations for the word *вода* (water) and some others by means of the tool Collocations of the NoSketch Engine system using 7 association measures. It is important to stress that the experiments were conducted using representative corpora, with large amount of the resulting collocations being under study. The data on the measure precision allows to establish to some degree that in cases when collocation extraction is not used for some special purposes such measures as *MI.l-og_f, log-Dice*, and *minimum sensitivity* should be used. No measure is ideal, which is why various options of their integration are desirable and useful. And we propose a number of parameters that allow to rank collocates in an integrated list, namely, an average rank, a normalised rank and an optimised rank.

Key words: collocation extraction, association measures, evaluation, average rank, normalised rank, optimised medium rank

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ КОЛЛОКАЦИЙ: СРАВНЕНИЕ И ИНТЕГРАЦИЯ МЕР АССОЦИАЦИИ

Захаров В. П. (v.zakharov@spbu.ru)

Санкт-Петербургский государственный
университет, Санкт-Петербург, Россия

1. Introduction: Collocations and Collocability

According to the theory of signs, the language system is based on two main types of relations between language units: paradigmatic and syntagmatic ones. Syntagmatic relations are based on linear speech (text) nature and occur on all language levels, though they are mostly studied in the framework of lexicology and syntax. When constructing a sentence, the choice of words is determined not only by their denotative and significative meanings, but rather it depends on the surrounding words which they are grammatically and semantically related to. The combinatory ability of language units, collocability, is one of the linguistic laws. However, these laws have not been studied in depth yet.

Let's speak now about the notion of collocation. There are different approaches to this term. Sometimes a collocation is meant as a synonym of a word combination, sometimes it is a special type of a set phrase. S. Evert suggests the following definition: "A collocation is a word combination whose semantic and syntactic properties can't be fully predicted on the basis of information about its constituents and which therefore should be added to the dictionary (lexicon)" [Evert 2004: 17]. However, explanatory dictionaries do not always consistently reflect the information about collocations. Perhaps, this is not even possible for a print dictionary of the whole language. Besides, the boundary between free and set phrases is quite vague. There are many set phrases whose meaning is equal to the sum of the meanings of their constituents, despite the fact that such phrases function as a single unit, with the stability rather than idiomatic nature being the main feature. A threshold of stability should be chosen to range them, above which a word combination can be called a set phrase. This approach assumes a probabilistic nature of collocations. Many modern authors and most of corpus linguists understand collocations as statistically determined set phrases. In this case, not only phrasemes and idioms, but also multiword terms, named entities (real-world objects, such as persons, locations, organisations, products, etc.,) and other types of free combinations could be regarded as set phrases.

The above approach is the basic point of our paper which is aimed at evaluation of various statistical methods of automatic collocation extraction.

2. Statistical background: association measures

Nowadays, there are several ways to calculate the degree of coherence of parts of a collocation. It is only natural to assume that one of the ways to identify the stability of a word combination is the frequency of their co-occurrence. The co-occurrence, in its turn, is associated with the frequency of individual components of the collocation. A whole number of formulae have been created (or borrowed from other sciences) to integrate different factors that determine the association between the collocation components. Usually, such formulae are called association measures. Most of them are based on the frequency comparison for pairs of words obtained by means of an actual corpus with relative frequencies taken from a hypothetical corpus where all words from the actual one are randomly located. It is carried out to identify the statistically significant fluctuations between the observed and expected frequencies [Church, Hanks 1991, Dunning 1993, Sinclair 1996, Evert., Krenn 2001].

There are different measures based on the calculation of the degree of words nearness in a text. P. Pecina provides 82 measures, describes their mathematical foundations including their formulae and key references [Pecina, 2009: 44–45, 48] (see also dissertation of S. Evert [2004]).

One should not forget also that words which tend to collocate with each other cannot be found in a random order in any case, as there exist grammar rules which imply that "the language system is a probabilistic one and it is a grammatical probability that word frequency shows in a text" [Halliday 1991: 31]. There are also methods that take into account the syntactic nature of collocations. B. Daille claims that the linguistic knowledge drastically improves the quality of stochastic systems [Daille 1994: 192]. One of the methods to take syntax in account are the so-called word sketches, which are lists of statistical collocations, each one for each syntactic relation [Kilgarriff, Tugwell 2001; Kilgarriff, Tugwell 2002]. These syntax-based collocations are described in detail by V. Seretan [2011: 59–101].

But in this paper, the grammatical probability is not taken into consideration, only the statistical one.

3. Association measures functionality

Lexical association measures being applied to a key word (node) occurrence and context statistics extracted from the corpus for all collocation candidates result in their association scores. A list of the candidates ranked according to their association scores is the desired result of the entire process. The top of the list are word combinations that are assumed to have the greatest association with each other and, consequently, be the most probable collocation candidates.

In general, all of them take into consideration the frequency of joint occurrence of a key word (node) and its collocate, thus answering the question how random the association force is between the collocates. But proper formulae are different, and they demonstrate different association strengths for the same collocations, which is why collocation ranks obtained by different measures do not coincide. It seems interesting and useful to try to reveal the functionality of different measures. It is known, too, that some measures bring similar results and others are significantly different [Křen 2006: 246–247].

The research on and evaluation of various association measures has been done for quite a long time and has been quite intensive [Dunning 1993; Evert, Krenn 2001; Braslavskij, Sokolov 2006; Pecina, 2009]. It is known that *t-score* extracts most frequent collocations. *Log-likelihood* was eventually preferred for its good behaviour on all corpus sizes and also for promoting less frequent candidates. On the contrary, the MI measure allows to reveal low-frequency multiword terms and proper names. Furthermore, it should be noted that the raw frequency of pairs was also found to be a good indicator of termhood, but it has the disadvantage of not being able to identify rare terms [Daille, 1994: 172–173].

Besides, association score depends on the type of the units (lemmas or word forms) whose statistics are used for the calculations. Sometimes collocation extraction by statistical measures has to be done on the word form level rather than on the lemma level. The analysis described in [Zakharov, Khokhlova 2014: 340] has shown that in some cases word form collocations overwhelmingly have significantly bigger value for all association measures.

The very number of the calculated collocates and the value of the association measure are also dependent on the “window” between the node and the collocate that has been chosen for the calculations.

4. Collocation extraction: integration of different association measures

The experiments were conducted on the basis of the Araneum corpora of Russian (http://unesco.uniba.sk), with the access provided through the NoSketch Engine. These corpora belong to the family of web corpora being created by the wacky technology [Benko 2014; Benko, Zakharov 2016]. We used 2 corpora, Russicum Russicum Minus (120 mln. tokens), Russicum Russicum Maius (1,20 bln.).Both consist of texts downloaded from the .ru domain sites. The access to corpora is provided through the NoSketch Engine [Rychlý 2007].

We extracted collocations for the words *вода* (water), *враг* (enemy) and *рыба* (fish) by means of the tool *Collocations* of the NoSketch Engine system using 7 association measures: *T-score*, *MI*, *MI3*, *log likelihood*, *minimum sensitivity*, *logDice* and *MI.log_f* [Statistics Used in Sketch Engine]. These measures are popular in many other systems, too. The major part of experiments was conducted on Russicum Russicum Minus corpus. This article provides the data obtained for the collocation window (–3, +1).

The result us represented by a list of collocates (collocations) organized for each of the 7 above association measures ranged according to the association score in the form of a table (see an example for the query *вода* (water) in Table 1). The number of collocates for each query was 200.

Table 1. List of collocates for вода (water) extracted by means of MI.log_f measure (a fragment)

Collocates	Co-occurrence count	Candidate count	MI.log_f score
Сточный (sewer)	12,479	13,791	100,505
Питьевой (drinkable)	11,288	14,006	97,878
Грунтовый (ground)	8,672	11,598	94,132
Кипяченый (boiled)	3,635	4,502	86,016
Горячий (hot)	20,665	102,240	84,393
Минеральный (mineral)	9,409	45,044	78,146
Холодный (cold)	15,172	102,915	77,386
Талый (melt)	1,863	2,701	77,295
Проточный (flowing)	2,602	5,125	77,246
Дистиллированный (distilled)	1,517	1,849	77,021
Пресный (fresh)	2,124	3,883	76,077

Collocates	Co-occurrence count	Candidate count	MI.log_f score
Подсоленной (salty)	1,082	1,211	74,331
Дождевой (rain)	2,279	5,476	73,727
Литр (litre)	9,275	63,939	73,217
Околоплодных (delivery)	767	806	71,279
Теплый (warm)	13,473	136,681	70,910
Кипеть (boil)	2,637	9,274	70,789
…………………	……	……	……

A rank has been assigned to every collocate in a table for each measure according to the score of the appropriate measure.

The next part of the study is aimed at developing methods for the integrated use of different measures of association. We used 7 collocation lists obtained in the first experiment. The ranged lists of collocates extracted by the 7 above association measures were processed in the following manner. Meaningless collocations with punctuation marks were removed. Due to errors of lemmatization, some collocates were presented in several different word forms. For such cases, non-lemmatized word forms of the same word were united into a single unit, with the highest association value being chosen. “Clean” lists of collocates were obtained as a result. Then, 7 tables (with 100 collocates in each) were merged into a new one in such a way so as to the collocates that were obtained through several measures were merged into a single line of the summary table, with their rank for each measure being provided (Table 2). When a collocate was not available among the first hundred collocates for appropriate measure it was not ranked.

Table 2. Summary table of collocates for *вода* (water)

Collocates	T-score	MI	MI3	log likelihood	min. sensitivity	log-Dice	MI.log_f
Сточный (sewer)	5	25	1	2	5	4	1
Питьевой (drinkable)	7	39	2	4	7	6	2
Грунтовый (ground)	13	53	4	7	13	10	3
…	…	…	…	…	…	…	…
Отвод (drainage)	60		35	37	64	50	29
Родниковый (sping)		78	70				30
Туалетный (cologne)	73		37	45	75	57	31
Обеззараживание (decontaminated)			47	69			32
…	…	…	…	…	…	…	…

It is clear that the same collocations with the word *вода* in the ranked list of different measures have different rank. Then a question arises: what is the rank of a certain collocation in such merged list, or, in other words, what single rank should be assigned for each collocation.

The following hypotheses were made:

1) the more the number of the measures in this combined list that identified a relevant collocate, the stronger the collocability of a given collocation;
2) the less the sum of the ranks or the average rank for a relevant collocate, the stronger the collocability, and, consequently, this sum can be regarded as the coefficient of the "value" of this collocation: the lesser sum makes a given collocation more "valuable" (potentially stronger);
3) if both conditions are observed then the "value" of a given collocation is even higher, which is why we introduce a normalised rank.

As a result, the following indicators have been added to Table 3:

1) the *number of association measures* that have "calculated" a given collocate (within 100 "cleaned" lines for each measure);
2) the *average rank* of the collocate: the sum of all ranks divided by the value "the number of association measures";
3) the *normalised rank* of the collocate (Table 3).

Table 3. Summary table of collocates for *вода* (water)

Collocates	T-score	MI	MI3	log likelihood	min. sensitivity	log-Dice	MI.log_f	Number of measures	Avera-ge rank	Nor-malised rank
Сточный (sewer)	5	25	1	2	5	4	1	7	6.14	6.14
Питьевой (drinkable)	7	39	2	4	7	6	2	7	9.57	9.57
Грунтовый (ground)	13	53	4	7	13	10	3	7	14.71	14.71
…	…	…	…	…	…	…	…			
Отвод (drainage)	60	0	35	37	64	50	29	6	45.83	51.33
Родниковый (spring)		78	70				30	3	59.33	103.23
Туалетный (cologne)	73	0	37	45	75	57	31	6	53.00	59.36
Обеззараживание (decontamination)	0	0	47	69	0	0	32	3	49.33	85.83
…	…	…	…	…	…	…	…			

The normalised rank is derived from the average rank multiplied by the coefficient that is in inverse proportion to the number of the association measures that have "calculated" a given collocate (NB: the less the rank, the more valuable the collocation is).

The coefficient for the normalised rank is calculated by the following formula:

$$log_2(1+7/n),$$

where n is the number of the successful measures for this collocate.

It is safe to say that the average and the normalised ranks "objectify" (integrate) the functionality of various association measures.

It should be noted that the less the rank the stronger (in theory) is the strength of the association between the collocation components. However, the rank is determined based on association measure score, which is why it is our task to correlate the ranks, i.e. the association strength, with some truth criterion.

5. Evaluation

Usually, comparison to some “gold standard“ or expert evaluation are used to evaluate the results of automated systems. When methods of collocation extraction are evaluated both options appear to be problematic. There is no „gold standard“ that would fully or significantly cover the set phrases. We could try to build it *ad hoc* for selected key words based on various dictionaries, but, due to the incomplete nature of dictionaries, the quality would be doubtful. As to expert evaluation, it is very expensive, taking into account time and human resources. The majority of automatic methods of word combination identification use large amount of data and result in large collocation lists. It would be prudent here to mention the volume of the sample evaluated. Expert method of evaluation usually covers only a small part of data due to its labour intensity. Unfortunately, the quality of automated methods is often evaluated based on the examples taken from the top units of ranked lists, and from a small number of the resulting collocates [Seretan 2011: 70].

In this paper, we have used expert evaluation, which means that each of 247 collocations of the summary list was marked either as a set phrase or not. According to the evaluation results, 86 collocations out of 247 were marked as true.

Further, we calculated the number of true collocations for each measure (within first 100 “cleaned” lines) (Table 4, the last line). The resulting number can be interpreted as the precision indicator (in percentage) for the upper part of the ranked list. This is also the indicator of the recall (or quasi-recall), i.e. how many collocations out of the potential 86 were obtained using this measure.

However, it is not only the number of the true collocations extracted using each measure that is important: the rank of the relevant collocation is significant, too. This is why it would be prudent to introduce a weight of true collocations for each measure taking into account the place of the collocates in a sorted table.

In order to evaluate the efficiency of each of the association measures the Kharin-Ashmanov method, which evaluates the relevance of the information retrieval results, was used [Ashmanov et al., 1997].

Based on the expert evaluation of the extracted collocates and their place in the ranked list with regard to each association measure, a characteristic set was formed. A characteristic set means the number of the true collocations obtained with different numbers of collocates from the ranked list (precision value).

According [Ashmanov et al., 1997], we select characteristic sets that contain 5 elements—the precision values for the first 10, 30, 50, 70 and 100 collocates from the top of the list (Table 4).

Table 4. Distribution of the number of true collocations for each measure

Ranks	T-score	MI	MI3	log likelihood	min. sensitivity	log-Dice	MI.log_f
1–10	0	5	4	2	5	6	8
1–30	6	10	11	9	11	9	15
1–50	8	18	18	14	13	15	23
1–70	10	28	21	22	14	19	26
1–100	17	39	33	26	21	22	31

A weight is assigned to each element of the characteristic set (5, 4, 3, 2, and 1, respectively). Each element is "weighed": each of 5 precision values is multiplied by the its weight and divided by 15 (the sum of the weights). The sum of the weighed elements is the resulting precision of the characteristic set.

Here is an example for the *MI* measure that has 5 true collocates in the top ten candidates (precision is 0.5), 10 true collocates in the top thirty (precision is 0.33), 18 in the top fifty (0.36), 28 in the top seventy (70), and 39 in the top hundred (0.39). Then, the resulting (average) precision will be equal to 0,5*5/15 + 0,33*4/15 + 0,36*3/15 + 0,4*2/15 + 0.39*1/15 = 0,167 + 0,088 + 0,072 + 0,053 + 0,026 = 0,406.

The values of the precision (let's call it *normalised precision*) calculated like that for all seven measures are given below (Table 5).

Table 5. Normalised precision values for association measures

	t-score	MI	MI3	log likelihood	min. sensitivity	log-Dice	MI.log_f
Number of true collocations	17	39	33	26	21	22	31
Normalised precision	0.115	**0.406**	0.366	0.262	0.357	**0.391**	**0.562**
Place	7	**2**	4	6	5	**3**	**1**

So, the best measure is *MI.l-og_f.* It is also shown that, for example, the precision of the *log-likelihood* measure (that has 26 true collocations) is lower than that of the *min. sensitivity* measure (that has 21 true collocation). The *MI* measure that has come second shall also be highlighted, since it is quite peculiar, despite its high performance. Here is the fragment from its collocate list (Table 5).

Table 6. Some collocates of *вода* (water) extracted by *MI* measure

	Co-occurrence count	Candidate count	MI
Омагниченная (magnetic)	6	5	11.581
Бахмут (bakhmut)	14	13	11.425
Бутилированная (bottled)	36	35	11.359
Мицеллярная (micellar)	10	10	11.318
Умягченная (soft)	5	5	11.318

This list can be continued: *деаэрируемая (deaerated), азотно-радоновая (nitrogen and radon), юрско-девонская (Jurassic & Devonian), подзоленная (ashen-gray), водородонасыщенная (reach in hydrogen).* On the one hand, this measure often extracts actual multi-word terms. On the other hand, it fails to identify or place into the "tail" of the list of well-known collocations. Erroneous spelling (*ки-пяченой, еесентуки*), proper names ("*Сент-Ронанские воды*"), nonce words, foreign words, words in Latin characters, etc. seem to be *MI* measure collocates, too. It should be noted that a lot of such words occur in large corpora. Of course such a "noise" in the corpus data can influence the results.

This is why we set a minimum limits for number of collocates in a corpus (the Sketch Engine has such parameters) to cut rare collocations, so that words with a frequency below a limit were not been included in the calculation.

New experiment with such a limitation and with other words gave the next results (Table 7, 8).

Table 7. Normalised precision values for association measures for *враг* (enemy) (with limitation of number of collocates)

	t-score	MI	MI3	log likelihood	min. sensitivity	log-Dice	MI.log_f
Number of true collocations	21	32	31	29	33	33	30
Normalised precision	0,266	0,505	0,362	0,373	**0,506**	**0,613**	**0,532**
Place	7	4	6	5	**3**	**1**	**2**

Table 8. Normalised precision values for association measures for *рыба* (fish) (with limitation of number of collocates)

	t-score	MI	MI3	log likelihood	min. sensitivity	log-Dice	MI.log_f
Number of true collocations	29	32	57	50	63	62	69
Normalised precision	0,229	0,340	0,572	0,495	**0,753**	**0,771**	**0,820**
Place	7	6	4	5	**3**	**2**	**1**

We see that MI measure in this case went down and that the best measures with respect to an expert evaluation are *MI.log_f, log-Dice* and *minimum sensitivity*. Similar results were obtained on the Russicum Russicum Maius corpus. So, we could conclude that efficiency of measures doesn't depend on corpus volume, at least, it is true for homogeneous web corpora.

Having obtained objective evaluation of the efficiency of individual measures, now we can introduce another rank indicator, which we will call the *optimised average rank*. This indicator is calculated taking into account the preference of the measures.

It is calculated as follows: all products of non-zero ranks multiplied by the coefficient of the measure significance are summed up and are divided into the number of measures used for a given collocate.

We suggest to set the measure significance coefficients, with their normalised precision taken into account (see Table 5, 7, 8 and also data from other experiments) as follows: *MI.log_f*—0.4, *logDice*—0.5, *min. sensitivity*—0.6, MI—0.7, MI3—0.8, *log-likelihood*—0.9, *T-score*—1.0. Of course, this is only preliminary ranking. The procedure of calculating normalised precision for association measures should be repeated on more words from different frequency belts.

As a result, the rank of the collocations extracted by more efficient measures is reduced, and the relevant collocate in the summary table goes up. See the example in Table 9.

Table 9. Optimised average rank for individual collocations

No.	Collocate	Average rank	Optimised average rank
1.	Поверхностный (surface)	**81.5**	59.8
2.	Крещенский (baptismal)	82.0	**36.0**
3.	Обычный (usual)	**61.0**	34.1
4.	Газированный (sparkling)	63.0	**27.9**
5.	Качество (quality)	**24.6**	19.5
6.	Урез (encroachment line)	29.0	**14.5**
7.	Соленый (salt)	**49.7**	37.2
8.	Паводковый (flood)	52.5	**22.5**

If you compare the collocates with even and odd numbers by pairs (*поверхностный* ‘*surface*’ vs. *крещенский* ‘*baptismal*’, *обычный* ‘*usual*’ vs. *газированный* ‘*sparkling*’, *качество* ‘*quality*’ vs. *урез* ‘*encroachment line*’, *соленый* ‘*salt*’ vs. *паводковый* ‘*flood*’), then it is clear that the latter, having collocations with *water* as the node, still have a bit higher average rank than the former (which means that their average stability obtained through the integration of all the association measures is a bit lower). However, as per our suggestion, following the optimisation, the latter will have a lower rank and go up in the ranked list (once again: the higher the rank in this list the higher the collocability degree).

6. Conclusion

To sum it up, the experiments have produced important results that characterise the efficiency of individual association measures. It is important to stress that the experiments were conducted using representative corpora, with large amount of the resulting collocations being under study.

We offer a method of assessing the effectiveness of statistical association measures. The data on the normalised precision allows to establish to some degree that in cases when collocation extraction is not used for some special purposes such measures as *MI.l-og_f, log-Dice*, and *minimum sensitivity* should be used. The *MI* measure is critical when rare multi-word terms are needed to be extracted.

Conversely, no measure is ideal, which is why various options of their integration are desirable and useful. And we propose a number of parameters that allow to rank collocates in such combined list. Merging several lists of collocates obtained by different measures into one improves the efficiency of statistical tools in total. We offer several options that allow to assess "the quality" of collocations in the combined list.

7. Further work

Further research will be as follows:

1. Develop the programming tool that allows to make a single list of collocates with all the necessary parameters and calculate integrated ranks.
2. Study how the efficiency of the association measures is associated with the width of the range (to the left and to the right of the key word) within which collocates are selected, and estimate the degree of such efficiency.
3. Identify the inter-relation between "syntagmatic" and "paradigmatic" collocates on the one hand and "idiomatic" and "statistical" on the other hand within the same search results, and identify the dependence of such inter-relation on the width of the window.
4. Do research with data from dictionaries used as the gold standard.

Acknowledgement

This work was partly supported by the grant of the Russian Foundation for Humanities (research project No. 16-04-12019).

References

1. *Ashmanov I., Grigoryev S., Gusev V., Kharin N., Shabanov V.* (1997), Using Statistical Method for Intelligent Computer-Based Text Processing [Primenenie statisticheskih metodov dlja intellektual'noj komp'juternoj obrabotki tekstov] / The Proceedings of the Dialog'97 International Seminar on Computational Linguistics and Its Applications, pp. 33–37.
2. *Benko V.* (2014), Aranea: Yet another Family of (Comparable) Web Corpora, Text, Speech, and Dialogue. 17th International Conference, TSD 2014 Brno, Czech Republic, September 8–12, 2014, Proceedings, Ed. P. Sojka et al., pp. 247–256.
3. *Benko V., Zakharov V. P.* (2016), Very large Russian corpora : new opportunities and new challenges. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2006"], Vol. 15 (22). pp. 79–93.
4. *Braslavskij P., Sokolov J.* (2006), Comparison of four methods of automation extraction of two-word terms from text [Sravnenie četyreh metodov avtomatičeskogo izvlečenija dvuhslovnyh terminov iz teksta], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue

2006" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2006"], Bekasovo, pp. 88–94.

5. *Church K., Hanks P.* (1991), Word Association Norms, Mutual Information and Lexicography, Computational Linguistics, Vol 16:1, pp. 22–29.
6. *Daille B.* (1994), Mixed approach for the automatic extraction of terminology: lexical statistics and linguistic filters [Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques], PhD thesis, Université Paris 7.
7. *Dunning T.* (1993), Accurate Methods for the Statistics of Surprise and Coincidence, Computational Linguistics Vol. 19, Issue 1, pp. 61–74.
8. *Evert S., Krenn B.* (2001), Methods for the Qualitative Evaluation of Lexical Association Measures, ACL Proceedings of 39th Annual Meeting, Toulouse, France, pp. 188–195.
9. *Evert S.* (2004), The Statistics of Word Cooccurences Word Pairs and Collocations, PhD thesis. Institut für Maschinelle Sprachverarbeitung (IMS). Stuttgart.
10. *Halliday M.* (1991), Current Ideas in Systemic Practice and Theory. London.
11. *Kilgarriff A., Tugwell D.* (2002), Sketching words, Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins, M. H. Correard (Ed.), Euralex, August, pp. 125–137.
12. *Kilgarriff A., Tugwell D.* (2001), WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography, Proc. Workshop COLLOCATION: Computational Extraction, Analysis and Exploitation, 39th ACL & 10th EACL, Toulouse, France, pp. 32–38.
13. *Křen M.* (2006), Collocation Measures and the Czech Language: Comparison on the Czech National Corpus data [Kolokační míry a čeština: srovnání na datech Českého národního korpusu], Kolokace, Praha, pp. 223–248.
14. *Pecina P.* (2009), Lexical Association Measures. Collocation Extraction, Prague.
15. *Rychlý P.* (2007), Manatee/Bonito — A Modular Corpus Manager, 1st Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Masaryk University, pp. 65–70.
16. *Seretan V.* (2011), Syntax-based Collocation extraction. Text, Speech and Language, Springer Science.
17. *Sinclair J.* (1996), The Search for Units of Meaning, Textus, IX, pp. 75–106.
18. Statistics Used in Sketch Engine. URL: https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine/ (Last access 03.02.2017).
19. *Zakharov V., Khokhlova M.* (2014), Syntagmatic Relations in Russian Corpora and Dictionaries, Pragmantax II. The Present State of Linguistics and its Sub-Disciplines [Zum aktuellen Stand der Linguistik und ihren Teildisziplinen], Frankfurt a.M., Peter Lang, pp. 333–344.