СОВМЕСТНАЯ ВСТРЕЧАЕМОСТЬ СЛОВ: ОПЫТ КЛАССИФИКАЦИИ

Паперно Д. А. (denis.paperno@unitn.it)

Università degli studi di Trento, Тренто, Италия

Ройтберг А. М. (cvi@yandex.ru),
Хачко Д. В. (mordol@lpm.org.ru),
Ройтберг М. А. (mroytberg@lpm.org.ru),

ИМПБ РАН, Пущино, Россия;
НИУ Высшая школа экономики, Москва, Россия

Ключевые слова: сочетаемость, коллокации, тематические модели, повторы

BREEDS OF COOCCURRENCE: AN ATTEMPT AT CLASSIFICATION

Paperno D. A. (denis.paperno@unitn.it)

Università degli studi di Trento, Trento, Italy

Roytberg A. M. (cvi@yandex.ru),
Khachko D. V. (mordol@lpm.org.ru),
Roytberg M. A. (mroytberg@lpm.org.ru)

IMPB RAS, Pushchino, Russia; RSU HSE, Moscow, Russia

The paper proposes a substantial classification of collocates (pairs of words that tend to cooccur) along with heuristics that can help to attibute a word pair to a proper type automatically.

The best studied type is frequent phrases, which includes idioms, lexicographic collocations, and syntactic selection. Pairs of this type are known to occur at a short distance and can be singled out by choosing a narrow window for collecting cooccurrence data.

The next most salient type is topically related pairs. These can be identified by considering word frequencies in individual documents, as in the well-known distributional topic models.

The third type is pairs that occur in repeated text fragments such as popular quotes of standard legal formulae. The characteristic feature of these is that the fragment contains several aligned words that are repeated in the same sequence. Such pairs are normally filtered out for most practical purposes, but filtering is usually applied only to exact repeats; we propose a method of capturing inexact repetition.

Hypothetically one could also expect to find a forth type, collocate pairs linked by an intrinsic semantic relation or a long-distance syntactic relation; such a link would guarantee co-occurrence at a certain relatively restricted range of distances, a range narrower than in case of a purely topical connection, but not so narrow as in repeats. However we do not find many cases of this sort in the preliminary empirical study.

Key words: cooccurrence, collocations, topic models, repeats

1. Introduction

Word cooccurrence has innumerous applications in computational linguistics. Much of the early research on co-occurrence focused on lexicographic tasks, using association measures to form lists of candidate multiword expressions to be included in dictionaries (e.g. [Smadja 1993]). Cooccurrence data have further usages in improving parsing algorithms (as in e.g. Yoon et al. [2001]) or as cue on a word's semantics; practical uses of such cues lead to the development of the distributional semantic models (DSMs). Applications of DSMs, range from semantic similarity recognition to word sense induction [Tamir and Rapp 2003], to word sense disambiguation [Mitrofanova et al. 2008], to entailment, to predicting association norms etc. A better understanding of the nature of co-occurrence, to which we aim to contribute here, could help improve many of these computational linguistic models.

Standardly, collocation extraction is based on corpus statistics. Collocations (in the broad, non-lexicographic sense) are word pairs that co-occur more often than expected from the frequencies of individual words and the statistical model adopted [Herbst 1996, Nesselhauf, 2004]. This approach was pioneered by [Firth 1957] and his followers [Halliday 1961] and [Sinclair 1991, Sinclair, Carter 2004]. See [Evert 2004] for methods of collocation extraction and an extensive literature overview.

The linguistic nature of statistically associated collocates varies, and our paper attempts at an exhaustive if coarse-grained classification. One of the best studied collocation types is idiomatic expressions: phrases whose meaning is not reduced to the meanings of constituent parts, such as *set forth*or *real estate*. Another subtype is lexicographic collocations, i.e. phrases with a more or less compositional meaning that are the default expression of a certain complex idea established in usage (e.g. *hard rain*, *strong tea*). Unlike idioms, lexicographic collocations allow for certain variation of phrase components while keeping the meaning largely intact (*strong sweet tea*vs. **real expensive estate*). approaches to lexicographic collocations include Mel'cuk [1998], Fillmore and Key [1988, 1992], and others.

There are cases of statistical association beyond multiword expressions such as idioms discussed above. For example, it has been noticed that members of a semantic field tend to co-occur (this is in fact a subclass of the "clustering" type, see below). However, no systematic classification of collocations by linguistic nature has been proposed to date. This paper fills this gap, proposing such a classification along with heuristics that allow for automatic attribution of collocations to one or another class, which we apply to the Brown corpus [Francis and Kucera, 1964].

This paper reports work in progress; further directions are outlined in section 4.

2. Materials and methods

2.1. Corpus

We conducted a preliminary quantitative analysis of the small but well-balanced Brown corpus. The Brown corpus contains 500 text fragments of approximately 2,000 words each. We lemmatized and retagged the corpus using FreeLing software [http://nlp.lsi.upc.edu/freeling/], which attributed the corpus's 1,010,058 words to 48,153 distinct lemmas.

2.2. Collocation extraction

As a measure of association significance we use the standardized deviation of pair frequency from the expected mean, known as the Z score:

$$Z(w_1, w_2, d) = \frac{f_C - E}{\sqrt{E}}$$

where f_c is the absolute cooccurrence frequency of w_1, w_2, at a given distance d. $E = f_1 \times (f_2/N)$ is the maximum-likelihood estimate of the mean and dispersion for co-occurrence frequency of w_1, w_2 at distance d, assuming the independence of occurrence of w_1 and w_2, where f_1 is the frequency of w_1, f_2 is the frequency of w_2 , N is the corpus size. Although the choice of association measure does affect the ranking of pairs, we note that the set of top pairs remains comparable when switching between measures. For example, among the word pairs that occur at least 3 times in the Brown corpus, in the lists of top 10K pairs the Z and the t scores share 71.8% of pairs, the Z score and PMI 95.8%; t score and PMI 77.5% of pairs. We use Z as our primary measure.

We identify associated pairs, or collocations in the broad sense, as pairs with Z score over 5. The first word in each pair was selected among the top 5K most frequent content words (verbs, nouns, adjectives, adverbs, or numerals). In addition, we used frequency thresholds; the quantitative results below include pairs that occur at least 3, 5, or 7 times. We also discuss the data obtained at the threshold of 5 in more detail.

3. Results

3.1. Classification of collocations

3.1.1. General

There are three groups of collocations with different linguistic nature and distribution.

- «phrases» are multiword expressions with an immediate syntactic relation between words;
- «repeats» are members of (nearly) identical text fragments such as legal formulae, see 3.1.3;
- «clustering» collocations are conditioned by corpus heterogeneity, see 3.1.4.

We conjecture that there is no substantial class of word association beyond these three.

Repeat-based collocations are filtered by entropy (3.1.3), clustering-based collocations are those whose Z score falls below 5 when calculated as described in 3.1.4. The heuristics proposed here are preliminary. For example, one could use methods other than entropy to identify repeats, or rely on paragraphs or other units rather than documents for detecting clustering effects.

The residual collocations are mostly phrasal collocations, which lie within the distance range of 3–5. A small number of statistically associated pairs do not seem to stand in a meaningful relation, and are not attributed to any of the three groups by heuristics. We believe that these data are mostly explained as noise, see discussion below.

3.1.2. Phrasal collocations

This class includes syntactically related associated words, in particular, the idioms and lexicographic collocations mentioned above. Table 1 contains some examples.

Table 1. Examples of phrasal collocations

№	Word 1	Word 2	Distance	Frequency	Z score
1	real	estate	1	24	231.30
2	urethane	foam	1	12	374.66
3	arc	voltage	2	5	160.49
4	great	deal	2	43	138.93
5	play	role	3	10	45.45
6	write	letter	3	12	33.03

Quite a few lexical collocations appear at a range of distances rather than a fixed distance, cf.:

Table 2. *Make clear* at distances 1–5

Distance	Word1	Word 2	Frequency	Z score
1	make	clear	13	17.459
2	make	clear	16	21.653
3	make	clear	6	7.673
4	make	clear	1	0.683
5	make	clear	0	−0.725

3.1.3. Repeats

Repetition of text fragments or clichés is quite common in corpora of naturally occurring text. This repetition raises association scores between all words in the repeat regardless of the distance or syntactic relation between them. Repeats can result from direct copying, as it often happens when one electronic document is created on the basis of another, but they can also stem from natural formulaic expressions. For example, Document H08 (Rhode Island Governor's Proclamations) of the Brown Corpus contains seven proclamations, each ending with the following: «In testimony whereof I have hereunto set my hand and caused the seal of the State to be affixed this 17th day of May in the year of Our Lord one thousand nine hundred and sixty-one». The formula repeats 7 times almost exactly (only the dates vary), boosting the association between all words in it. For instance, the pair *affix this* gets the Z score of 25.16. Another inexact repeat with more variable elements is *between N p.m. and K a.m.*, raising the association between *p.m.* and *a.m.* to 134.2 (Z score), with 4 occurrences at distance 3.

To identify a repeat, we take all occurrences of a pair at a given distance, and calculate the entropy of each position in those contexts. We rely on the heuristic that a small average entropy for all positions for the window containing the given word pair, all positions between them, and 10 positions on each side, indicate a repeat. For each position we calculate entropy as

$$\mathrm{E} = -\sum\nolimits_{\mathrm{i}} \left(\mathrm{P(i)} \times \ln\left(\mathrm{P(i)}\right)\right)$$

where i ranges over words, and P(i) is the probability of word i in the given position.

We take the threshold for average entropy across all positions to be 0.8. In contrast to existing approaches, the entropy-based method allows us to identify even inexact repetition.

3.1.4. Clustering

Corpus structure affects statistical word association. If some part of a corpus has higher frequencies of words *x* and *y* than the rest, we also expect it to contain more pairs of *x* and *y*. So even independent occurrence of *x* and *y* in the subcorpus may lead to statistical association given the overall frequencies of *x* and *y*. Let's show the role of corpus heterogeneity by an example.

Verbs *tell* and *think* are quite frequent, these lemmas occur 766 and 1,044 times respectively in the Brown corpus, with relative frequencies $p_1 = 766 \div 1{,}010{,}058 = 0.076\%$ and $p_2 = 1{,}044 \div 1{,}010{,}058 = 0.104\%$. Assuming that these verbs are distributed independently (the null hypothesis for any word pair), the probability of finding *tell* and *think* at any given fixed distance is $p_1 \times p_2$, with $p_1 \times p_2 \times C$ expected occurrences of the pair, where C is corpus size. So we can expect the pair *tell*, *think* to appear roughly once at each distance ($0.076\% \times 0.104\% \times 1{,}010{,}058 = 0.798$).

Now imagine that both *tell* and *think* occur exclusively in fiction, which contributes about a quarter of the Brown corpus, and are not attested elsewhere. In this case the fiction corpus should contain all the pairs of these words, and the expected number can be obtained by multiplying the frequencies of *tell* and *think* in the fiction subcorpus by its size (about 252K words), i.e. $(766 \div 252{,}000) \times (1{,}044 \div 252{,}000) \times 252{,}000 = 3.17$. In fact, the lemmas *tell* and *think* are attested 7 times at distance 10, which corresponds to the Z score of 6.98 (assuming independence of occurrence and expected frequency of 0.798) or 2.15 (assuming that both lemmas occur only in fiction). As one can see, taking into account corpus heterogeneity can lead to significantly different association measures.

(Of course, for *tell* and *think* both models are crude idealizations. The truth is in the middle: for the 7 occurrences of *tell* and *think*, the Z score based on actual corpus heterogeneity is 4.69, almost exactly between 6.98 and 2.15.)

In practice, for almost all lemmas w_1, w_2 there are several rather than two subcorpora characterized by different frequencies of w_1 and w_2. For the purpose of this paper, we treated each document as a potentially distinct thematic subcorpus. This assumption is harmless: if in fact documents form blocks characterized by even word frequency distributions, the sum of expected frequencies for all documents will give, on average, a correct estimate of the expected frequency for the whole block. As the expected overall frequency of a pair in the corpus, we take the sum of expected frequencies for all documents:

$$E = -\sum_D \left(f_{1(D)} \times \left(f_{2(D)} / N_D\right)\right)$$

where D ranges over all documents, N_D is the size of D, $f_{1(D)}$, $f_{2(D)}$ are the frequencies of w_1, w_2 in D. This calculation is valid regardless of how diverse document sizes are. Adjusted association scores such as Z can then be calculated on the basis of this corrected E. If such an adjusted score of a collocation is low, then the collocation owes its initially high association measure solely to clustering of both words in the same documents.

Of course we do not imply that the property of two lexemes to occur in the same texts is irrelevant. To the contrary, it reveals an association through features of genre, style, or topic; this includes sameness of semantic field. What we want to emphasize is that association by clustering is substantially different from other types of collocation, and should be separated for practical applications. For instance, extraction of lexicographic collocations might be improved by disregarding the effects of clustering, while for the study of topic structure only clustering effects are relevant, but not other types of co-occurrence.

We also note that for clustering collocations, the specific numeric value of association is an artifact of corpus composition. Indeed, it was quite an arbitrary decision on behalf of the creators of the Brown corpus to dedicate just a quarter of its size to fiction, as opposed to a bigger or a smaller part. But it is the fraction of texts of each topic and genre that determines how high the association measure will be for words characteristic of that topic or genre.

Corpus heterogeneity creates quite many associated pairs. Examples include pairs *member—church*, *student—college*, *state—federal*, which are not spread across the whole corpus but are clustered in a small set of documents. For *member—church* the basic Z score is 14.35, but it drops to 4.19 when adjusted word frequencies in individual documents; *student—college* has Z scores of 18.05 and 3.26, *state—federal* 15.73 and 4.44, respectively. All of these pairs illustrate a primarily topical relation between words.

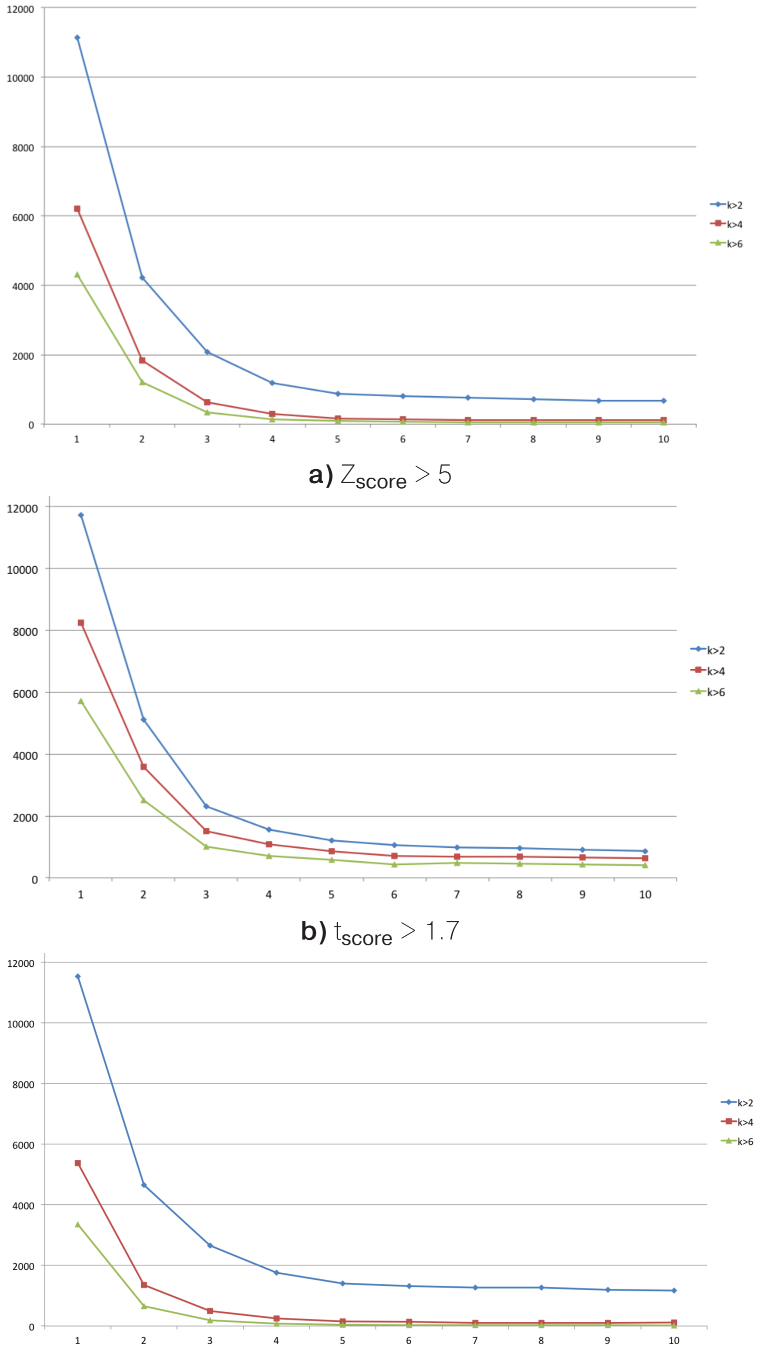
3.2. Quantitative observations

3.2.1. Distance and collocation type.

Figure 1.a–c shows the dependence of the number of collocations found on distance, cf. 2.2. Both content and function words are included. Distribution shape is stable across association measures (a–c).

As the graphs show, at distance of 5 the number of collocations stabilizes, while short distances (1 and 2) contribute many more pairs. Between 3 and 5 the number of collocations decreases relatively slowly. This pattern agrees with the standard assumption that collocations are mostly found at distances up to 3–5 words. Our own informal observations on the lists of collocations agree with this assumption.

The pattern is the same for all three thresholds. In what follows we use only the 5 threshold.



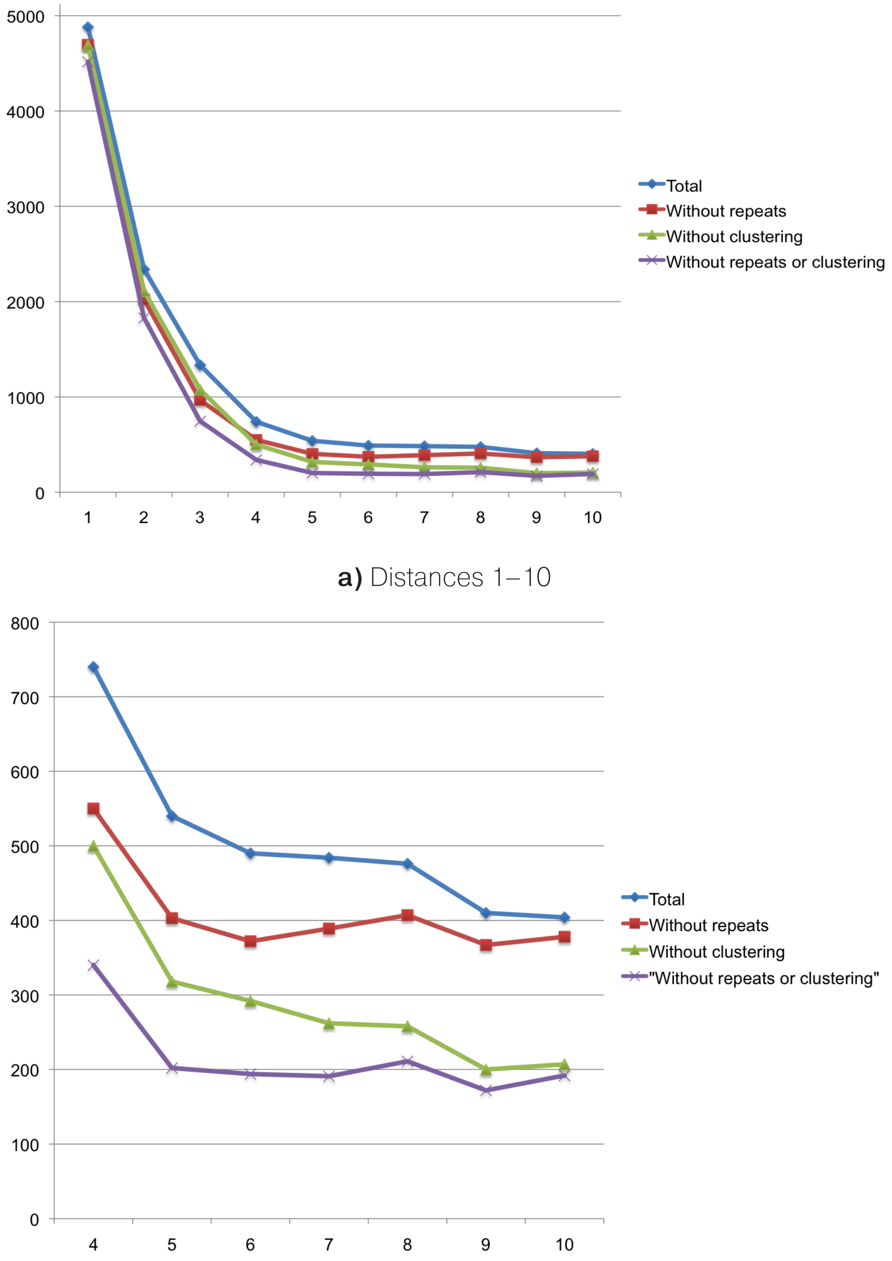
a) $Z_{score} > 5$

b) $t_{score} > 1.7$

c) PMI > 0.72

Fig. 1. Number of collocations by distance and frequency threshold

3.2.2. Distribution of individual types of collocations



a) Distances 1–10

b) Distances 4–10

Fig. 2. Number of collocations of different classes; only content words considered

As we see from graph 2, on distances (greater than 5) the majority of pairs belong to the "clustering" collocations and repeats. Most of the rest, by our qualitative observations, is statistical noise. We deal with pairs of words which exhibit no meaningful relation(*say — New York*, *number — eye, so — work*). Raising the frequency threshold essentially eliminates those "remote" collocations. Some of the "remote" collocations are mostly due to the clustering effect, but were not shelled out by the formal criterion because of random statistical fluctuation on top of the clustering effect. One such example is the pair *optimal — state*: its Z score at distance 8 drops from 32.6 to 6.7 when taking corpus heterogeneity into account, and the pair also occurs at distances of 6 and 9 (Z drops from 6.39 to 0.8) and 7 (from 19.48 to 3.75).

A priori, there could be more types of remote collocations: two words could be related by a relation of syntactic or discourse nature at a greater distance. We could expect that a pronoun is anteceded by a coreferent name (anaphora), that after mentioning the evil an author is likely to talk about the good (associative relation), or that discourse markers would tend to occur in a certain order (*although…still, on the one hand...on the other hand*), one sentence or in different sentences.All of these cases are genuine word-word relations at remote distances, as opposed to links mediated by a popular quote or by the text topic. , the analysis of the Brown data reveals almost no examples of this kind. The only exceptions are pairs of markers *first...then*, *only... also*.

4. Discussion

While individual factors of statistical association have been noticed previously (cf. [Evert 2004]), this paper is the first attempt at a substantial classification of associated pairs by the main underlying factor. The novel tentative conclusion of this paper is that the three types discussed here exhaust all the statistical collocations. One practical consequence, briefly discussed below, is that the number of different types of collocations could be a useful characteristic of a corpus.

The classification proposed here can help improve any practical applications of lexical association measures, from collocation extraction to refining distributional semantic models that build semantic vectors based on association measures [Turney, Pantel 2010]. For small distances, it will be interesting to evaluate how much filtering repeats and clustering helps identify true lexicographic collocations.

Automatic classification of collocations that we implemented can also serve as a basis for qualitative assessment of natural language corpora. In particular, a corpus of identical size should be more valuable for most applications if it has fewer repeats. Perhaps even more significant could be the number of "clustering" collocations. Indeed, if each of those pairs points to a particular topic or genre represented by a distinct subcorpus, then abundance of such topical pairs, other things being equal, tells us that the corpus is diverse and balanced. The balancing effect arises because if a certain topic takes up a disproportionately large part of the corpus, the word pairs that correspond to the topic get lesser weight. In an analogous but more balanced corpus the same topic will contribute more statistical collocations thanks to a greater degree of clustering. We leave the development of a specific procedure of corpus evaluation to future research.

References

1. *Francis W. N., Kucera H.* (1964), Department of Linguistics, Brown University, Providence, Rhode Island, USA. http://icame.uib.no/brown/bcm.html.
2. *Tamir R., Rapp R.* (2003), Mining the Web to Discover the Meanings of an Ambiguous Word. IEEE International Conference on Data Mining — ICDM, pp. 645–648.
3. *Bell E. J. L.* (2007), Collocation Statistical Analysis Tool: An evaluation of the effectiveness of extracting domain phrases via collocation. B.Sc. Dissertation, Lancaster University.
4. *Evert S.* (2004), The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgar.
5. *Fillmore Ch., Kay P., O'Connor M.* (1988), Regularity and idiomaticity in grammatical constructions: The case of LET ALONE, Language, Vol. 64, pp. 501–518.
6. *Firth J. R.* (1957) Modes of Meaning, Papers in Linguistics 1934–51, pp. 190–215, Oxford University Press.
7. *Halliday M. A. K.* (1961), Categories of the Theory of Grammar, Word 17, pp. 241–92.
8. *Herbst T.* (1996) What Are Collocations: Sandy Beaches or False Teeth? English Studies No. 4, pp. 379–393.
9. *Manning C., Schutze H.* (1999) Foundations of Statistical Natural Language Processing, MIT Press. Cambridge.
10. *Mel'cuk, I.* (1998). Collocations and Lexical Functions. Cowie, A. P. (ed.), Phraseology. Theory, Practice and Applications, Oxford University Press, pp. 23–53, Oxford.
11. *Mitrofanova O. A., Belik V. V., Kadina V. V.* (2008), Corpus Analysis of Selectional Preferences of Frequent words in Russian [Korpusnoe issledovanie sochetaemostnyh predpochtenij chastotnyh leksem russkogo jazyka], Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog 2008". Moscow.
12. *Nesselhauf N.* (2004) Collocations in a Learner Corpus, Amsterdam/Philadelphia, Benjamins.
13. *Padró L., Stanilovsky E.* (2012) FreeLing 3.0: Towards Wider Multilinguality. Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey.
14. *Sinclair J.* (1991), Corpus, Concordance, Collocation, Oxford University Press, Oxford.
15. *Sinclair, J., Carter, R.* (2004) Trust the Text. Language, Corpus and Discourse, Routledge, London/New York.
16. *Zaharov V. P., Hohlova M. V.* (2010) Study of Effectiveness of Statistical Measures for Collocation Extraction on Russian Texts, Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference Dialog 2010.