

# Survey of Word Co-occurrence Measures for Collocation Detection

Olga Kolesnikova

Instituto Politécnico Nacional, Superior School of Computing Sciences (ESCOM),  
Mexico City, Mexico

kolesolga@gmail.com

**Abstract.** This paper presents a detailed survey of word co-occurrence measures used in natural language processing. Word co-occurrence information is vital for accurate computational text treatment, it is important to distinguish words which can combine freely with other words from other words whose preferences to generate phrases are restricted. The latter words together with their typical co-occurring companions are called collocations. To detect collocations, many word co-occurrence measures, also called association measures, are used to determine a high degree of cohesion between words in collocations as opposed to a low degree of cohesion in free word combinations. We describe such association measures grouping them in classes depending on approaches and mathematical models used to formalize word co-occurrence.

**Keywords.** Word co-occurrence measure, association measure, collocation, statistical language model, rule-based language model, hybrid approach to model word co-occurrence.

## 1 Introduction

Knowledge of lexical co-occurrence and of lexical relation accounts for the extent to which the choice of a word in a text is stipulated by its surrounding words without taking into account syntactic and/or semantic reasons [39]. Such knowledge is very important in many tasks of natural language processing: text analysis and generation, knowledge extraction, opinion mining, text summarization, question answering, machine translation, polarity identification, information retrieval, among others.

For instance, in the text generation task, one is interested in construction of not only grammatically correct utterances but also of those that sound natural. In order to achieve this, the system must know restrictions on usage of a particular word,

that is, its combinability with other words in an utterance, or its collocational preferences.

Basically, word combinations can be divided into two big classes depending on word collocational choices. These two classes are free word combinations and restricted word combinations, also termed collocations.

Usually, collocations are defined as characteristic and frequently recurrent combinations [10, 11] of (commonly) two linguistic elements which have a direct syntactic relationship [40] but whose co-occurrence in texts cannot be explained only by grammatical rules [7].

One of the elements of a collocation is called a base or node and is autosemantic, that is, it can be interpreted even if it is not in the context of the collocation [13, 14]. The other element called collocator or collocate is semantically dependent on the base, has a more opaque meaning, and can only be interpreted with reference to the collocation, that is, it is synsemantic [14].

## 2 Strategies of Measuring Word Co-occurrence for Collocation Detection

In natural language processing research, there have been developed basically three strategies for automatic learning restrictions on word usage: statistical, rule-based, and hybrid strategies. Generally speaking, a computer system is expected to analyze a machine-readable text or a corpus defined as a collection of machine readable texts. The system must be able to extract combinations in which words are syntactically related and determine to what extent the appearance of one word in a phrase depends on

the occurrence of another word or words. Such feature is called cohesion or association.

Within the statistical strategy, which is most common in language processing and lexical co-occurrence research, in order to calculate a measure of word association within collocations, a formal model of word co-occurrence should be designed or selected from the existing statistical models.

An important advantage of statistical models is that they use raw corpora where a selected language unit (word, type or lemma, phrase, sentence, document) is viewed as a data item. Statistical modeling is attractive since conclusions are derived out of data in a way that seems much more objective in comparison with linguistic interpretations and theories based on introspection and intuition of experts in linguistics. Moreover, statistical analysis in general is language-independent.

However, statistical methods work under certain assumptions, for instance, that data items "obey" certain well-studied distributions: normal, binomial,  $\chi^2$ , or other distributions. We do not know actually how real life linguistic data is distributed, but in any case, our mathematical constructs can be justified by the golden principle of pragmatics: it works therefore it is true. Another problem of statistical methods is that they require large corpora, otherwise estimations of frequencies and probabilities of word co-occurrences become imprecise and untrustworthy. Besides, if a collocation has a very low frequency of occurrence, it can hardly be detected.

In order to combat the above mentioned problems associated with the statistical methods, the rule-based strategy was put forward. Methods developed within this trend allow detecting low frequency phrases and do not rely on a very large collection of data.

On the other hand, rule-based techniques usually depend on language and lack flexibility. The latter characteristic harms the detection of those collocations which permit syntactic variation. Also, making hand-crafted rules is time consuming. Moreover, such rules have limited coverage and will hardly discover new collocations appearing in language.

In an effort to overcome the disadvantages of both strategies mentioned above and to take

advantage of positive aspects of the same strategies, the hybrid methods have been proposed. Such methods use rules to extract candidate phrases and then apply statistical methods to improve the obtained results.

In this paper, we consider in detail the three strategies—statistical, rule-based, hybrid—on the task of detection of collocations. Reviewing each strategy, we describe various methods developed in state of the art works within the strategy, discuss their degree of effectiveness, and give examples.

### **3 Statistical Strategy to Measure Word Co-occurrence**

Within the statistical methodology, candidate word combinations are identified based on calculation of a predetermined association measure in  $n$ -grams extracted from a corpus. Usually,  $n$ -grams are word combinations of a chosen syntactic pattern, e.g., adjective+noun or verb+preposition depending on the preferred structural type. In order to do this, the chosen corpus is lemmatized; words are tagged with their respective parts of speech (POS-tagged). Also, the corpus can be parsed. Evidently, this preprocessing is language-dependent. Another feature used for  $n$ -grams extraction is window size, typically from 1 to 5 words.

After  $n$ -grams are extracted, the association strength between their constituents is computed according to some statistical metric. As we have mentioned previously, such metrics used in the process of extracting word combinations are termed word co-occurrence measures or association measures because they compute the degree of association between the components in a phrase.

In this section we consider the association measures used to detect two-word collocations, i.e., bigrams, which is a very common case. Besides, the association measures for bigrams can be extended to combinations of three or more words.

Pecina [25] gives a comprehensive list of 82 association measures used to detect two-word collocations. To calculate the association measures, it is common to take into account frequencies of occurrence of each word in a bigram

... součástí trhu, vznikl obratem <b>černý trh</b> s plyšovými medvídky a ...
zabránit přísnu drog na <b>domácí černý trh</b> v hodnotě 32 milionu .....
stejnými jednotlivci i <b>kompletní černý trh</b> . Jinými slovy, byla by .....
... pomáhali pašování cigaret na <b>černý trh</b> do východního Německa ...
....nájemních práv <b>nezaručený černý trh</b> . Libor Dellin, člen .....

... miliónů dolarů. Ovlivňuje nějak negativně tento <b>černý trh</b> naše hospodářství? Je to pouze ztráta na daních
.. <b>Maltské liry lze nakoupit pouze ve směnárnách, černý trh</b> s valutami neexistuje. Na Maltě je v porovnání s
operoval i ženu. A přece má, jak se říká na Arbatu, <b>černý trh</b> něco do sebe. Je - li hlad nejlepší kuchař, je .....
.. přestal. <b>V patách za krizi vstoupil do Bělehradu černý trh</b> , pašování a zvýšená kriminalita. Překupníci ...
.....z toho obviněni. <b>ídí gangy</b> , které kontrolují <b>černý trh</b> a okrádají cizince. Oba byli zbaveni funkcí a byl

**Fig. 1.** Example of a left immediate context (top) and empirical context (bottom) of the collocation *černý trh* (black market)

$xy$  (a sequence of two words, the word  $x$  and the word  $y$ ), the frequency of the bigram, its immediate context, and its empirical context.

The words  $x$  and  $y$  are viewed as types or lexical items, i.e., words as they are encountered in a lexicon. Their realizations are various grammatical forms found in a text. Commonly, frequencies are estimated for types, and we view frequencies here in this manner. However, the theory termed lexical priming states that word associations are characteristic of different forms of lexical items, so a particular wordform may have its own collocations typical for it and not typical for another form of the same lexical item [16]. Therefore, lexical priming aims at a more fine-grained classification of word associations which is its strong side. On the other hand, such approach increases the complexity of analysis, since lexical items may have a very big number of grammatical forms. Also, frequencies of each wordform may be not high enough and thus not sufficient for statistical tests to work accurately, as the total frequency of a type is distributed through the whole range of its numerous forms thus obtaining low frequencies for each form of the type.

Speaking about the context of a bigram  $xy$ , we mentioned above that in calculating association measures the immediate and empirical contexts are used [26]. The immediate context of a bigram is word(s) immediately preceding or following the bigram. The empirical context of a word sequence is open class words occurring within a specified context window. Open class words include nouns, verbs, adjective, and adverbs.

Figure 1 gives an example of the immediate and empirical contexts of the Czech bigram *černý trh* (*black market*) from [25]. In this example, the left immediate context includes one word, and the empirical context contains all words of the utterance where the given bigram is used taken without this bigram.

In the formulas of association measures which we discuss in detail in what follows, the notation from the contingency table is used. The term *contingency table* was first used by Karl Pearson in 1904, and such table is a certain manner of considering the occurrence of two words symbolized as  $x$  and  $y$ .

The contingency table presented in Table 1 contains observed frequencies for a bigram  $xy$ . In this table, the following notation is used:  $f(xy)$  is the frequency or the number of occurrences of the bigram  $xy$  in a corpus;  $\bar{x}$  stands for any word except  $x$ ,  $\bar{y}$  stands for any word except  $y$ ,  $*$  stands for any word;  $N$  is a total number of bigrams in a corpus.

In fact,  $N$  can be interpreted differently depending on the task of the application being developed or on the objective of research, so

**Table 1.** Contingency table of co-occurrence frequencies of a bigram  $xy$  and its constituent words  $x$  and  $y$

$a = f(xy)$	$b = f(x\bar{y})$	$f(x*)$
$c = f(\bar{x}y)$	$d = f(\bar{x}\bar{y})$	$f(\bar{x}*)$
$f(*y)$	$f(*\bar{y})$	$N$

generally speaking,  $N$  is the number of language units in the corpus chosen for consideration. Such language units can be tokens, types,  $n$ -grams of tokens or types, sentences, documents, etc. The choice of a language unit depends on the granularity of semantic analysis. In this article, we interpret  $N$  as the total number of bigrams in a corpus.

The bigrams are obtained following the paths in trees of syntactic dependencies or constituents resulting from parsing, so the words in a bigram are syntactically related. This procedure filters out irrelevant combinations of words which do not comprise a phrase with meaningful semantic interpretation.

Frequencies  $f(x^*)$ ,  $f(\bar{x}^*)$ ,  $f(*y)$ ,  $f(*\bar{y})$  in the contingency table are called marginal totals or simply marginal frequencies, and  $f(xy)$ ,  $f(x\bar{y})$ ,  $f(\bar{x}y)$ ,  $f(\bar{x}\bar{y})$  are called joint frequencies.

In formulas, the contingency table cells are sometimes referred to as  $f_{ij}$ . Statistical tests of independence also work with contingency tables of expected frequencies  $\hat{f}(xy)$  defined as

$$\hat{f}(xy) = \frac{f(x^*)f(*y)}{N}.$$

In the contingency table, the following holds:

$$\begin{aligned} f(x^*) &= f(xy) + f(x\bar{y}), \\ f(\bar{x}^*) &= f(\bar{x}y) + f(\bar{x}\bar{y}), \\ f(*y) &= f(xy) + f(\bar{x}y), \\ f(*\bar{y}) &= f(x\bar{y}) + f(\bar{x}\bar{y}), \\ N = f(N) &= f(x^*) + f(\bar{x}^*) \\ &= f(*y) + f(*\bar{y}). \end{aligned}$$

In some formulas of association measures, the concept of probability is used. The probability of finding a word  $x$  in a corpus  $P(x^*)$  is calculated according to the formula

$$P(x^*) = \frac{f(x^*)}{N},$$

where  $f(x^*)$  is the frequency of  $x$  in a corpus and  $N$  is the corpus size.

Also, in the formulas of some association measures, the context is represented by the following notation:

- $C_w$  is the empirical context of  $w$  ( $w$  stands for any word),
- $C_{xy}$  is the empirical context of  $xy$ ,
- $C_{xy}^l$  is the left immediate context of  $xy$ ,
- $C_{xy}^r$  is the right immediate context of  $xy$ .

We remind the reader that some examples of the immediate and empirical context are given in Figure 1.

## 4 Typology of Statistical Association Measures

Evert [8] proposes a comprehensive classification of statistical association measures. They can be calculated using the UCS toolkit, software written in Perl of the same author (available at <http://www.stefan-evert.de/Software.html>). Evert defined the four approaches within the statistical strategy, and within each approach, a number of types of association measures. The classification is as follows:

**Approach 1.** The methods within the first approach measure the significance of association between the words  $x$  and  $y$  in a bigram  $xy$ . They quantify the amount of evidence that the observed bigram  $xy$  provides against a null hypotheses of independence of the words  $x$  and  $y$  in this bigram, i.e.,  $P(xy) = P(x^*)P(*y)$ , or against the null hypothesis of homogeneity of the columns in the contingency table for this bigram (for details on the null hypothesis of homogeneity see section 2.2.4 in Evert 2005). The methods in this approach are the following:

- Likelihood measures which compute the probability of the observed contingency table (multinomial-likelihood, binomial-likelihood, Poisson-likelihood, the Poisson-Stirling approximation, and hypergeometric-likelihood);
- Exact statistical hypothesis tests which compute the significance of the observed data (binomial test, Poisson test, Fisher's exact test);
- Asymptotic statistical hypothesis tests used to compute a test statistic (z-score, Yates' continuity correction, t-score which compares the observed co-occurrence frequency  $f(xy)$  and the expected co-occurrence frequency  $\hat{f}(xy)$  as random variates, Pearson's chi-

squared test, Dunning's log-likelihood which is a likelihood ratio test).

**Approach 2.** The methods within the second approach measure the degree of association of the words  $x$  and  $y$  in a bigram  $xy$  by estimating one of the coefficients of association strength from the observed data. This class includes measures of two types:

- Point estimates, usually, maximum-likelihood estimates (mutual information, odds ratio, relative risk, Liddell's difference of proportions, minimum sensitivity, geometric mean coefficient, Dice coefficient or mutual expectation, Jaccard coefficient);
- Conservative estimates based on confidence intervals obtained from a hypothesis test (a confidence-interval estimate for mutual information).

**Approach 3.** The techniques within the third approach measure the association strength of the words  $x$  and  $y$  in a bigram  $xy$  or, in other words, the non-homogeneity of the observed contingency table compared to the contingency table of expected frequencies. These methods take advantage of the concepts of entropy, cross-entropy, and mutual information borrowed from the information theory (pointwise mutual information, local mutual information, average mutual information).

**Approach 4.** The methods in the fourth approach use various heuristics to evaluate the degree of association between the components of a bigram  $xy$ . Usually, such methods apply modified versions of measures from the other three approaches or combine such measures (co-occurrence frequency, variants of mutual information, random selection).

Now using the notation given previously, in the following sections we consider various association measures used for automatic detection of collocations in natural language texts. In each section we indicate the approach to which the considered association measures belong, so we grouped these measures by their types following the typology presented above.

We did not put the methods in the numeric order of the approaches, rather we ordered them using

the criterion of complexity. First we describe some simple methods to estimate the association of words in a bigram  $xy$ , then we proceed to more complex formulas and techniques.

## 5 Simple Frequency-based Association Measures

In this section we discuss some simple measures, belonging to Approach 4, based on word frequency and probability used to detect collocations in a corpus.

In the simplest case, taking advantage of such property of collocations as recurrency (i.e., frequent usage in texts), we can count the number of occurrences of a bigram  $xy$  and estimate their joint probability  $P(xy)$ :

$$P(xy) = \frac{f(xy)}{N}.$$

If the bigram is used frequently, than it is probable that the two words are used together not by chance but comprise a collocation.

Also, to detect collocations, raw frequency of a bigram can be used instead of its probability. An example of this approach is the work of Shin and Nation [38] which presents most frequent collocations found by the authors in the spoken section of the British National Corpus (BNC). The article includes a list of 100 collocations ranked by their frequency in the BNC and in Table 2 we reproduce the upper part of this list which includes the most frequent word combinations.

The number of the bigram occurrences represented as the joint probability of two words occurring together  $P(xy)$  can be compared with probabilities of individual words  $P(x\bar{y})$  and  $P(\bar{x}y)$  in combinations with the words other than the one in the bigram  $xy$ .

A drawback of using the joint probability  $P(xy)$  is that this measure does not capture the direction of the relation between the word  $x$  and the word  $y$ . It means that the joint probability does not distinguish if  $x$  is more predictive of  $y$  or the other way round. That is, this measure (and the majority of other association measures) is bidirectional or symmetric [12]. In other words, the joint probability mixes two different probabilities: the conditional

**Table 2.** Most frequent collocations in the spoken section of the British National Corpus

Rank	Collocation	Collocation Frequency
1	<i>you know</i>	27348
2	<i>I think (that)</i>	25862
3	<i>a bit</i>	7766
4	<i>(always [155], never [87] used to {INF})</i>	7663
5	<i>as well</i>	5754
6	<i>a lot of {N}</i>	5750
7	<i>{No.} pounds</i>	5598
8	<i>thank you</i>	4789
9	<i>{No.} years</i>	4237
10	<i>in fact</i>	3009
11	<i>very much</i>	2818
12	<i>{No.} pounds</i>	2719
13	<i>talking about {sth}</i>	2489
14	<i>(about [91] {No.} percent (of sth [580], in sth [54], on sth [44], for sth [38]))</i>	2312
15	<i>I suppose (that)</i>	2281
16	<i>at the moment</i>	2176
17	<i>a little bit</i>	1935
18	<i>looking at {sth}</i>	1849
19	<i>this morning</i>	1846
20	<i>(not) any more</i>	1793

probability  $P(y|x)$  and the reverse conditional probability  $P(x|y)$  defined by the following equations:

$$P(y|x) = \frac{f(xy)}{f(xy) + f(x\bar{y})},$$

$$P(x|y) = \frac{f(x\bar{y})}{f(xy) + f(x\bar{y})}.$$

An example of the conditional probabilities approach is the works of Michelbacher, Evert, and Schütze [20, 21] where  $P(y|x)$  and  $P(x|y)$  are used for exploring adjective and/or noun collocates in a window of 10 words around node words in the British National Corpus.

The value of the conditional probability  $P(y|x)$  or the reverse conditional probability  $P(x|y)$  can also

be compared with the product of individual probabilities  $P(x*)$  and  $P(*y)$ . As a result of such comparison, it can be determined if the word  $y$  occurs independently of the word  $x$ , and in such case we get  $P(y|x) = P(x*)P(*y)$ , or the occurrence of  $y$  depends on the occurrence of  $x$ , that is,  $P(y|x) \neq P(x*)P(*y)$ . If we work with the reverse conditional probability, we can verify whether  $P(x|y) \neq P(x*)P(*y)$ . If the two words under consideration occur independently, then we deal with a free word combination, and a collocation otherwise.

## 6. Information-Theoretic Measures

The association measures in this section belong to Approach 3 and are based on such concepts as

mutual information and entropy. These metrics measure the mutual dependency between two words  $x$  and  $y$  which are constituents of a bigram  $xy$ .

### 6.1 Mutual Information (MI)

$MI$  is a well-known information-theoretic notion used to judge about dependence of two random variables. Its application as an association measure for collocation extraction was suggested by Church and Hanks [6].  $MI$  is an estimation of how much one word  $x$  tells about the other word  $y$  and it is computed according to the formula

$$MI = \log \frac{P(xy)}{P(x^*)P(*y)},$$

where the probabilities are calculated using data from the contingency table (see Table 1). We remind the reader that  $P(xy)$  is the joint probability of  $x$  and  $y$  co-occurrence:

$$P(xy) = \frac{f(xy)}{N},$$

and  $P(x)$  and  $P(y)$  are individual, or marginal, probabilities of  $x$  and  $y$ , respectively:

$$P(x^*) = \frac{f(x^*)}{N}, \quad P(*y) = \frac{f(*y)}{N}.$$

As it is seen from the formula for  $MI$ , the collocation hypothesis is expressed as the probability  $P(xy)$  actually observed in a corpus, and the null hypothesis suggests that  $x$  and  $y$  are independent, i.e. constitute a free word combination, therefore, the probability of co-occurrence  $P(xy)$  has the property

$$P(xy) = P(x^*)P(*y).$$

If  $MI = 0$ , the null hypothesis is proved, if  $MI > \theta$ , where  $\theta$  is a threshold estimated experimentally, then  $x$  and  $y$  are associated with the constituents of a collocation.

Strictly speaking, the metric we have just considered is pointwise mutual information. But in the NLP literature it is referred to as simply mutual information according to the tradition started in [6], since we are interested not in mutual information of

two random variables over their distribution, but rather in mutual information between two particular points.

Pecina and Schlesinger compared the effectiveness of 82 association measures given in their article [26] and demonstrated experimentally that pointwise mutual information works as the best association measure to identify collocations. However, this metric becomes problematic when data is sparse; it is also not accurate for low-frequency word combinations.

Two versions of mutual information are also applied to estimate the association strength of the words  $x$  and  $y$  in a bigram  $xy$ , these are average mutual information calculated according to the formula

$$\text{average-}MI = \sum_{i,j} f_{i,j} \cdot \log \frac{f_{i,j}}{\hat{f}_{i,j}}$$

and local mutual information for a given bigram  $xy$  which is estimated as follows:

$$\text{local-}MI = f(xy) \cdot \log \frac{f(xy)}{\hat{f}(xy)}.$$

In these formulas  $f_{i,j}$  is frequency in a cell of  $i \times j$  contingency table, in our example of  $2 \times 2$  table (see Table 1), the cells are  $f_{11}$ ,  $f_{12}$ ,  $f_{21}$ ,  $f_{22}$ .

### 6.2 Evaluation

Upon obtaining a list of collocation candidates, evaluation of the list must be done to check what candidate phrases are true collocations.

Evaluation can be manual or automatic. Results are presented in terms of conventional precision and recall. Given a finite set of word combinations, precision  $P$  is the number of word combinations correctly identified as collocations by the method under evaluation compared to all word combinations identified as collocations by the method; recall  $R$  is the same number of correctly identified collocations compared to all collocations in the dataset:

$$P = \frac{\text{#correctly identified as collocations}}{\text{#identified as collocations}},$$

$$R = \frac{\text{\#correctly identified as collocations}}{\text{\#collocations}}$$

Manual evaluation is fulfilled using three methods. The retrieved list of candidates is ordered and the first  $n$ -best candidates (with highest values of association measure applied in the extraction process) can be

- checked by a native speaker who has sufficient training in linguistics or by a professional lexicographer;
- compared against a dictionary, however, the evaluation results will depend on quality and coverage of the dictionary;
- evaluated using a hand-made gold standard (a list of collocations manually identified in a corpus).

Although manual evaluation is very accurate, it suffers certain limitations. If collocation candidates are evaluated by human experts, they may have disagreements on the status of some expressions. This is due to a lack of formality in the definitions of collocations as well as to the nature of this linguistic phenomenon since there are no clear-cut boundaries among various types of collocations and between collocations and free word combinations.

On the other hand, when evaluation is performed against a dictionary, the scope of work is restricted by the inventory of phrases in the selected dictionary. However, when collocations are extracted from very large corpora, the list of candidates is much bigger than the expressions found in the dictionary, therefore, a good portion of true collocations might be lost in the evaluation process.

Concerning a hand-made golden standard, the limit is time and financial resources because manual work is always costly in both senses. A very serious limitation of manual evaluation is the impossibility to estimate recall for very large lists of collocations candidates. It may seem that the problem can be solved with the data size reduction (to 50-200 samples), but association measures do not work well on small datasets.

To overcome the drawbacks of manual evaluation, automatic evaluation methods have been proposed. A well-known and widely used

method was developed by Evert and Krenn [9]. Instead of manually annotating only a small (in the sense of automatic language processing) number of  $n$ -best collocation candidates, Evert and Krenn suggested to compute precision and recall for several  $n$ -best samples of an arbitrary size comparing them against a golden standard of about 100 collocations (true positives, TPs). Then, in this case, precision is the proportion of TPs in the  $n$ -best list, and recall is the proportion of TPs in the base data that are also contained in the  $n$ -best list, the base data being an unordered list of all extracted collocation candidates.

## 7 Likelihood Measures

The association measures in this section belong to the first approach.

### 7.1 Log-Likelihood Ratio

The alternative terms for this measure found in literature are G-test and maximum likelihood statistical significant test. Log-likelihood ratio is computed with the formula

$$LR = -2 \sum_{ij} f_{ij} \log \frac{f_{ij}}{\hat{f}_{i,j}}.$$

Similar to the formula of average mutual information given in the previous subsection,  $f_{ij}$  is frequency in a cell of  $i \times j$  contingency table, in our example of  $2 \times 2$  table (see Table 1), the cells are  $f_{11}, f_{12}, f_{21}, f_{22}$ . The expected frequency  $\hat{f}_{i,j}$  is computed as if data items were independent, i.e., according to the formula given for the case of our contingency table

$$\hat{f}_{i,j} = \frac{f(\text{dataItem}_1)f(\text{dataItem}_2)}{N},$$

where  $\text{dataItem}$  is either  $x$ ,  $y$ ,  $\bar{x}$ , or  $\bar{y}$ , depending on what cell is considered.

Another option is to determine the squared log likelihood ratio according to the following formula:

$$\text{squared } LR = -2 \sum_{i,j} \frac{\log f_{ij}^2}{\hat{f}_{ij}}.$$

## 7.2 Multinomial Likelihood

This measure termed multinomial likelihood (*ML*) estimates the probability of the observed contingency table point hypothesis assuming the multinomial sampling distribution:

$$ML =$$

$$\frac{N!}{N^N} \cdot \frac{(\hat{f}(xy))^{f(xy)} \cdot (\hat{f}(x\bar{y}))^{f(x\bar{y})} \cdot (\hat{f}(\bar{x}y))^{f(\bar{x}y)} \cdot (\hat{f}(\bar{x}\bar{y}))^{f(\bar{x}\bar{y})}}{f(xy)! \cdot f(x\bar{y})! \cdot f(\bar{x}y)! \cdot f(\bar{x}\bar{y})!}.$$

## 7.3 Hypergeometric Likelihood

Another version is the hypergeometric likelihood (*HL*) computed under the general null hypothesis of independence  $P(xy) = P(x)P(y)$ :

$$HL = \frac{\binom{C_1}{f(xy)} \cdot \binom{C_2}{R_1 - f(xy)}}{\binom{N}{R_1}},$$

where  $R_1 = f_{11} + f_{12}$  and  $C_j = f_{1j} + f_{2j}$ .

## 7.4 Binomial Likelihood

Under the assumption of the binomial distribution, we can compute the total probability of all contingency tables for  $f(xy)$  according to the formula and obtain the binomial likelihood (*BL*):

$$BL = \binom{N}{f(xy)} \left( \frac{\hat{f}(xy)}{N} \right)^{f(xy)} \left( 1 - \frac{\hat{f}(xy)}{N} \right)^{N-f(xy)}.$$

## 7.5 Poisson Likelihood

If we replace the binomial distribution with the Poisson distribution, this will increase the computational efficiency and will provide results with a higher accuracy. In this case, the corresponding association measure is called Poisson likelihood (*PL*) and is calculated as follows:

$$PL = e^{-\hat{f}(xy)} \frac{(\hat{f}(xy))^{f(xy)}}{f(xy)!}.$$

## 7.6 Poisson-Stirling Approximation

This is another association measure calculated under the assumption of the Poisson distribution. If we take the negative logarithm of the Poisson likelihood and approximate the factorial  $f_{11}!$  in the formula for the Poisson likelihood given in the previous section with the Stirling formula  $n! \approx \sqrt{2\pi n} \left( \frac{n}{e} \right)^n$ , we obtain the following association measure called the Poisson-Stirling measure (*PSM*):

$$PSM = f(xy) \cdot (\log f(xy) - \log \hat{f}(xy) - 1).$$

## 8 Exact Statistical Hypothesis Tests

The association measures belonging to Approach 1 and estimated using the concept of likelihood as in the previous sections may suffer from very small values of probabilities which may be obtained in some cases. However, one can provide evidence concerning the independent occurrence of the words  $x$  and  $y$  in a bigram  $xy$  using what is called exact or strict hypothesis tests. These tests evaluate the probability of the null hypothesis  $H_0$  (stating that  $P(xy) = P(x)P(y)$ ) against the alternative hypothesis  $H_1$  of the words  $x$  and  $y$  being the collocation constituents. The probability, at which the decision to favor or reject the null hypothesis is made, is called the significance level of the test, and values of 10%, 5%, or 1% are usually used. The lower the value of the significance level, the stricter the test is.

For the binomial distribution, the following metric is used:

$$B = \sum_{k=f(xy)}^N \binom{N}{k} \left( \frac{\hat{f}(xy)}{N} \right)^k \left( 1 - \frac{\hat{f}(xy)}{N} \right)^{N-k}.$$

For the Poisson distribution, the following metric is used:

$$P = \sum_{k=f_{11}}^{\infty} e^{-\hat{f}(xy)} \frac{(\hat{f}(xy))^k}{k!}.$$

Another exact test can be developed using the hypergeometric likelihood function, this gives what is called the Fisher's exact test:

$$F = \sum_{k=f(xy)}^{\min\{R_1, C_1\}} \frac{\binom{C_1}{k} \cdot \binom{C_2}{R_1 - k}}{\binom{N}{R_1}} \\ = \frac{f(x^*)! f(\bar{x}^*)! f(*y)! f(*\bar{y})!}{N! f(xy)! f(x\bar{y})! f(\bar{x}y)! f(\bar{x}\bar{y})!}$$

where  $R_1 = f(xy) + f(x\bar{y})$  and  $C_j = f_{1j} + f_{2j}$ .

## 9. Asymptotic Statistical Hypothesis Tests

The methods in this group belong to Approach 1, they operate under the assumption of the normal distribution. The asymptotic theory is also called the large sample theory, and of course, natural language corpora are very large samples of linguistic data. Different from statistic tests of other groups which work on a finite data sample of size  $N$ , the asymptotic tests assume that the sample size grows infinitely and estimate test statistics for  $N \rightarrow \infty$ . Within the framework of the asymptotic theory, various test statistics have been developed and now we will consider the most common of them.

### 9.1 z-score

This is the simplest test statistic in this group of association measures. It is a simplification of computing the binomial measure approximating the discrete binomial distribution with the continuous normal distribution. Figure 2 borrowed from [8] shows how this approximation works.

The z-score is computed with the formula

$$z = \frac{f(xy) - \hat{f}(xy)}{\sqrt{\hat{f}(xy) \left(1 - (\hat{f}(xy)/N)\right)}}.$$

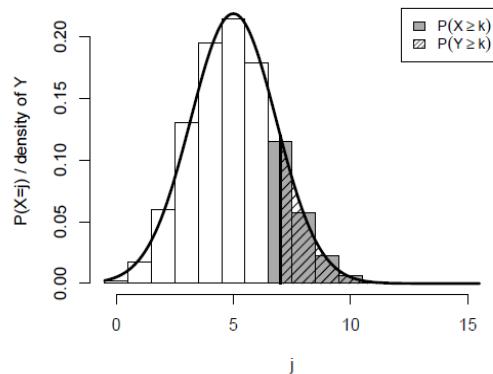


Fig. 2. Normal approximation  $Y$  to binomial distribution  $X$

### 9.2 Yates' Continuity Correction

The Yates' function is applied to the normal approximation to the binomial distribution. This correction is made to adjust observed frequencies towards the expected frequencies thus obtaining the corrected frequencies  $f_{ij}^{\text{corrected}}$  according to the formulas:

$$f_{ij}^{\text{corrected}} = f_{ij} - \frac{1}{2} \text{ if } f_{ij} > \hat{f}_{ij}, \\ f_{ij}^{\text{corrected}} = f_{ij} + \frac{1}{2} \text{ if } f_{ij} < \hat{f}_{ij}.$$

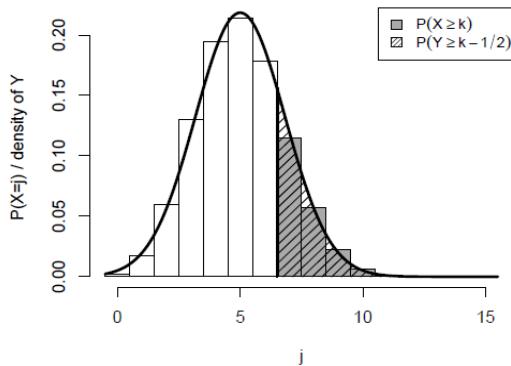
Figure 3 shows that Yates' continuity correction is a closer approximation to the binomial distribution, compare it with the Figure 2.

### 9.3 t-score

This metric, also called the Student's  $t$ -test, compares the observed co-occurrence frequency  $f(xy)$  and the expected co-occurrence frequency  $\hat{f}(xy)$  as random variates. The  $t$ -score is computed according to the formula

$$t = \frac{f(xy) - \hat{f}(xy)}{\sqrt{f(xy)(1 - (f(xy)/N))}}.$$

This test estimates whether the means of two groups of data are statistically different. Applying it to measuring the association of words in a bigram



**Fig. 3.** Normal approximation with Yates' correction

$xy$ , one thus compares the observed and expected frequencies. The above formula corresponding to the  $t$ -test measure is a ratio. The numerator is the difference between the observed and expected frequencies, and the denominator includes a measure of the variability of the observed frequency. The  $t$ -test requires the normal approximation and assumes that the mean and variance of the distribution are independent.

#### 9.4 Pearson's $\chi^2$ Test

The standard test for the independence of the rows and the columns in the contingency table is Pearson's  $\chi^2$  test. It is a two-sided association measure computed according to the formula

$$\chi^2 = \sum_{i,j} \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}}.$$

Another version of this test is based on the null hypothesis of homogeneity and is estimated according to the following formula:

$$\chi^2 = \frac{N(f(xy)f(\bar{x}\bar{y}) - f(x\bar{y})f(\bar{x}y))^2}{(f(xy) + f(x\bar{y}))(f(\bar{x}y) + f(\bar{x}\bar{y}))(f(xy) + f(\bar{x}y))(f(x\bar{y}) + f(\bar{x}\bar{y}))}.$$

#### 9.5 Dunning's Log-Likelihood

This test is a likelihood ratio test, and its metric  $LD$  is computed according to the formula

$$LD = -2\log \frac{L(f(xy), C_1, r) \cdot L(f(\bar{x}\bar{y}), r)}{L(f(xy), C_1, r_1) \cdot L(f(\bar{x}\bar{y}), r_2)},$$

$$\text{where } L(k, n, r) = r^k (1-r)^{n-k}, \quad r = \frac{f(xy) + f(\bar{x}\bar{y})}{N},$$

$$r_1 = \frac{f(xy)}{f(xy) + f(\bar{x}y)}, \quad r_2 = \frac{f(\bar{x}\bar{y})}{f(x\bar{y}) + f(\bar{x}\bar{y})}.$$

As it can be seen from the formula, this test compares the likelihood of two hypotheses about the words in a bigram  $xy$ , the first hypothesis is  $P(y|x) = P(y|\bar{x})$ , and the second hypothesis states that  $P(y|x) \neq P(y|\bar{x})$ .

In order to estimate the statistical significance of the calculated metric, it is multiplied by  $-2$ , and then one has to consult the  $\chi^2$  table at the degree of freedom equal to one.

## 10 Coefficients of Association Strength

The methods in this section belong to Approach 2. The techniques within this approach measure the degree of association between the words  $x$  and  $y$  in a bigram  $xy$  by estimating one of the coefficients of association strength from the observed data.

### 10.1 Odds Ratio

Odds ratio is the ratio of two probabilities, that is, the probability that a given event occurs and the probability that this event does not occur. This ratio is calculated as  $ad/bc$ , where  $a$ ,  $b$ ,  $c$ , and  $d$  are taken from the contingency table presented earlier (see Table 1).

The odds ratio is sometimes called the cross-product ratio because the numerator is based on multiplying the value in cell  $a$  times the value in cell  $d$ , whereas the denominator is the product of cell  $b$  and cell  $c$ . A line from cell  $a$  to cell  $d$  (for the numerator) and another from cell  $b$  to cell  $c$  (for the denominator) creates an  $\times$  or cross on the two-by-two table.

### 10.2 Relative Risk

This measure estimates the strength of association of the words  $x$  and  $y$  in a bigram  $xy$  according to the formula

$$\frac{\frac{a}{a+b}}{\frac{c}{c+d}},$$

where  $a, b, c$ , and  $d$  are taken from the contingency table presented earlier (see Table 1). The name of the measure is explained by the fact that this metric is commonly used in medical evaluations, in particular, epidemiology, to estimate the risk of having a disease related to the risk of being exposed to this disease. However, it also can be applied to quantifying the association of words in a word combination.

Relative risk is also called risk ratio because, in medical terms, it is the ratio of the risk of having a disease if exposed to it divided by the risk of having a disease being unexposed to it.

This ratio of probabilities can also be used in measuring the relation between the probability of a bigram  $xy$  being a collocation versus the probability of this bigram to be a free word combination.

### 10.3 Liddell's Difference of Proportions

This measure ( $LDP$ ) is the maximum likelihood estimation for the difference of proportions and is calculated according to the formula

$$LDP = \frac{f(xy)f(\bar{x}\bar{y}) - f(x\bar{y})f(\bar{x}y)}{f(*y)f(*\bar{y})}.$$

This metric has been applied to text statistics in [19], where one can find a detailed discussion of its advantages compared to the conditional exact test without randomization.

### 10.4 Minimum Sensitivity

Minimum sensitivity ( $MS$ ) is an effective measure of association of the words  $x$  and  $y$  in a bigram  $xy$  and has been used successfully in the collocation extraction task. This metric is calculated according to the formula

$$MS = \min \left\{ \frac{f(xy)}{f(x*)}, \frac{f(xy)}{f(*y)} \right\}.$$

In fact, what this measure does is comparing two conditional probabilities  $P(y|x)$  and  $P(x|y)$  and selecting the lesser value thus taking advantage of the notion of conditional probability.

However, [12] suggests that  $MS$  has not to be trusted without a proper consideration in spite of its good performance. The reason is that the value of this metric does not specify what association is taken into account, i.e., the association between  $x$  and  $y$  or the association between  $y$  and  $x$ . The author gives an example of the collocation *because of* and states that if the value  $MS = 0.2$  is obtained, this number does not reveal whether the 0.2 is  $\frac{f(xy)}{f(xy)+f(x\bar{y})} = P(y|x)$  or  $\frac{f(xy)}{f(xy)+f(\bar{x}y)} = P(x|y)$ , that is, whether it is  $P(of|because)$  or  $P(because|of)$ .

### 10.5 Geometric Mean Coefficient

This association measure is calculated according to the formula

$$gmean = \frac{f(xy)}{\sqrt{f(x*)f(*y)}},$$

or

$$gmean = \sqrt[N]{f(xy)}.$$

The geometric mean is similar to the arithmetic mean, however, they are different in the operation over which the average value is calculated. The arithmetic mean is applied when several numbers are added together to produce a total value. The arithmetic mean estimates the value that each of the summed quantities must have to produce the same total. That is, if all the summands had the same value, what this value would be to give the same total. Analogously, the geometric mean is applied to multiplication of several factors, and it estimates the value that each of the factors must have (the same value for all the factors) to produce the same value of the product.

**Table 3.** Collocates of *break* extracted using *t*-score, *MI*, and Dice coefficient

Collocate	<i>f(xy)</i>	<i>t</i> -score	Collocate	<i>f(xy)</i>	<i>MI</i>
the	11781	99.223	spell-wall	5	11.698
.	8545	83.897	deadlock	84	10.559
,	8020	80.169	hoodoo	3	10.430
be	6122	69.439	scapulum	3	10.324
and	5183	65.918	Yasa	7	10.266
to	5131	65.918	intervenient	4	10.224
a	3404	52.214	Preparedness	21	10.183
of	3382	49.851	stranglehold	18	10.177
down	2472	49.412	logjam	3	10.131
have	2813	48.891	irretrievably	12	10.043
in	2807	47.157	Andernesse	3	10.043
into	1856	42.469	irreparably	4	10.022
he	1811	39.434	Theif	37	9.994
up	1584	39.038	THIEf	4	9.902

Collocate	<i>f(xy)</i>	Dice coefficient
down	2472	0.0449
silence	327	0.0267
into	1856	0.0210
leg	304	0.0203
off	869	0.0201
barrier	207	0.0191
law	437	0.0174
up	1584	0.0158
heart	259	0.0155
neck	180	0.0148
news	236	0.0144
rule	292	0.0142
out	1141	0.0135
away from	202	0.0135
bone	151	0.0130

Applied to the contingency table (see Table 1), the geometric mean is equal to the square root of the heuristic  $MI^2$  measure defined by the following formula:

$$MI^2 = \log \frac{(f(xy))^2}{\hat{f}(xy)}.$$

Therefore, the geometric mean increases the influence of the co-occurrence frequency in the

numerator and avoids the overestimation for low-frequency bigrams.

#### 10.6 Dice Coefficient

This association measure (*D*) is calculated according to the formula

$$D = \frac{2f(xy)}{f(x*) + f(*y)}.$$

This coefficient is one of the most common association measures used to detect collocations; moreover, its performance happens to be higher than the performance of other association measures.

For example, Rychlý [37] experimented with various association measures including *t*-score, *MI*, *MI*<sup>3</sup> (defined as  $\log \frac{P^3(xy)}{P(x*)P(*y)}$ ), minimum sensitivity, *MI* log frequency (defined as *MI*  $\times$   $\log f(xy)$ ), and Dice coefficient with the objective to extract collocations from a corpus for lexicographic purposes.

The experiments showed that Dice coefficient outperformed the other association measures. Besides, collocations detected with Dice coefficient were relevant for a collocation dictionary. To show this, in Table 3 we reproduce the results of applying three measures, namely, *t*-score, *MI*, and Dice coefficient, to extract collocations of *break* in [37].

### 10.7 Jaccard Coefficient

The Jaccard coefficient (*J*) is monotonically related to the Dice coefficient and measures similarity in asymmetric information on binary and non-binary variables. It is commonly applied to measure similarity of two sets of data and is calculated as a ratio of the cardinality of the sets' intersection divided by the cardinality of the sets' union. It is also frequently used as a measure of association between two terms in information retrieval.

To estimate the relation between the words *x* and *y* in a bigram *xy*, the Jaccard coefficient is defined by the following formula:

$$J = \frac{a}{a + b + c},$$

where the values of *a*, *b*, and *c* are as given in the contingency table (see Table 1).

The Jaccard coefficient as well as the Dice coefficient are often called normalized matching coefficients because the way to assess the similarity of two terms is to count the total number of each combination in a contingency table as

presented in Table 1. Jaccard is similar to the cosine coefficient (*cos*) defined by the following formula:

$$\cos = \frac{a}{\sqrt{(a + b)} \times \sqrt{(a + c)}},$$

where the values of *a*, *b*, and *c* are as given in the contingency table (see Table 1) and on average, Jaccard and cosine have more than 80% agreement (for example, see the results of experiments in [5]).

### 10.8 Confidence-Interval Estimate for Mutual Information

Point estimates of association between words in a phrase operate well for words which have sufficiently high frequency, however, these metrics are not reliable when words or word combinations have few occurrences in a corpus. This fact results in a low performance of point estimates, for example, of mutual information, as shown in [9]. This issue can be resolved by using interval estimates from exact hypothesis tests which correct for random variation and evade overestimation.

The confidence-interval estimate for mutual information (*MIconf*) is defined as

$$MIconf = \log \min \left\{ \mu > 0 \left| e^{-\mu \hat{f}(xy)} \sum_{k=f(xy)}^{\infty} \frac{(\mu \hat{f}(xy))^k}{k!} \geq \alpha \right. \right\}.$$

## 11 Rule-Based and Hybrid Strategies to Measure Word Co-occurrence

Statistical methodology requires large collections of data, otherwise estimations of frequencies and probabilities of word co-occurrences become imprecise and untrustworthy.

However, the Zipf's Law asserts that the frequency of a word in a corpus is inversely proportional to its rank in the frequency table. Therefore, a great deal of words (from 40% to 60% of large corpora, according to [18]) are *hapax legomena*, i.e., they are used only once in a corpus.

**Table 4.** Candidate selection rules

Criteria	Rules
Frequency	<b>(Rule1)</b> Frequency heuristic: frequency $\geq 2$ for simplex words vs. frequency $\geq 1$ for NPs
Length	<b>(Rule2)</b> Length heuristic: up to length 3 for NPs in non-of-PP form vs. up to length 4 for NPs in of-PP form (e.g. <i>synchronous concurrent program</i> vs. <i>model of multiagent interaction</i> )
Alternation	<b>(Rule3)</b> of-PP form alternation (e.g. <i>number of sensor</i> = <i>sensor number</i> , <i>history of past encounter</i> = <i>past encounter history</i> ) <b>(Rule4)</b> Possessive alternation (e.g. <i>agent's goal</i> = <i>goal of agent</i> , <i>security's value</i> = <i>value of security</i> )
Extraction	<b>(Rule5)</b> Noun Phrase = $(NN NNS NNP NNPS JJ JJR JJS) \times (NN NNS NNP NNPS)$ (e.g. <i>complexity</i> , <i>effective algorithm</i> , <i>grid computing</i> , <i>distributed web-service discovery architecture</i> ) <b>(Rule6)</b> Simplex Word/NP <u>IN</u> Simplex Word/NP (e.g. <i>quality of service</i> , <i>sensitivity of VOIP traffic</i> ( <i>VOIP traffic extracted</i> ), <i>simplified instantiation of zebroid</i> ( <i>simplified instantiation extracted</i> ))

Low-frequency phenomenon also extends to collocations. Baldwin and Villavicencio [1] indicate that two-thirds of verb-particle constructions occur at most three times in the overall corpus.

An example of the rule-based method can be found in [17]. The authors use candidate selection rules for key phrase extraction from scientific articles. Key phrases are simplex noun or noun phrases that represent the key ideas of the document. Examples of rules are presented in Table 4.

On the other hand, rule-based techniques usually depend on language and lack flexibility. The latter characteristic harms the extraction of collocations which permit syntactic variation.

Also, making hand-crafted rules is time consuming. Moreover, such rules have limited coverage and will hardly discover new restricted word combinations appearing in language.

To combat these disadvantages, hybrid methods have been proposed. The latter use rules to extract candidate restricted constructions and apply statistical methods to improve the obtained results.

For example, in [15], machine learning is used together with simple patterns to identify functional expressions in Japanese. Their experiments show that the hybrid method doubles the coverage of previous approaches to resolving this issue, at the same time preserving high values of precision.

## 12 Conclusions

In this article we presented a detailed survey of word co-occurrence measures used in natural language processing. Such measures are called association measures and they are applied to determine the degree of word cohesiveness in phrases. If the value of such measure is high in a given word combination, the latter is called collocation. Collocations are different from free word combinations, in which the degree of cohesiveness is low. It is important in natural language processing to determine which word combination is a collocation since they must be treated in a way different from phrases in which words combine freely.

We described the association measures grouping them in classes depending on approaches and mathematical models used to formalize word co-occurrence. The three approaches were presented: statistical, rule-based, and hybrid approaches. Most association measures belong to the statistical approach in which there are many types distinguished: frequency-based measures, information-theoretic measures, likelihood measures, statistical hypothesis tests (exact and asymptotic), and coefficients of association strength. The measures are described indicating their formulas, basic principles of their definition, their advantages and disadvantages.

In the last decade, the area of NLP has changed radically with the progress of deep learning. There is a number of directions to improve the co-occurrence measure for collocation detection, such as the use of concepts and common sense knowledge [27, 32, 35, 36], as well as sentiment and emotion information in the text [29, 33]. In addition word2vec-like techniques can be used to improve the co-occurrence measure for collocation spotting. Word2vec is a vector-space language model learned using deep learning, which has shown good performance on text [2-4] and multimedia [28, 30, 31] analysis. Co-occurrence methods can be useful for personality detection [34] and textual entailment-based techniques [22-24].

## Acknowledgements

The author highly appreciates the support of Mexican Government which made it possible to complete this work: SNI-CONACYT, BEIFI-IPN, SIP-IPN: grants 20162064 and 20161958, and the EDI Program.

## References

1. **Baldwin, T. & Villavicencio, A. (2002).** Extracting the unextractable: A case study on verb particles. *Proceedings of the Sixth Conference on Computational Natural Language Learning (CoNLL2002)*, Vol. 20, pp. 99–105. DOI: 10.3115/1118853.1118854.
2. **Cambria, E., Poria, S., Bisio, F., Bajpai, R., & Chaturvedi, I. (2015).** The CLSA model: A novel framework for concept-level sentiment analysis. *International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2015, Lecture Notes in Computer Science*, Vol. 9042, Springer, pp. 3–22. DOI: 10.1007/978-3-319-18117-2\_1.
3. **Chikersal, P., Poria, S., & Cambria, E. (2015).** SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning. *SemEval-2015*, p. 647.
4. **Chikersal, P., Poria, S., Cambria, E., Gelbukh, A., & Siong, C.E. (2015).** Modelling public sentiment in Twitter: using linguistic patterns to enhance supervised learning. *International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2015, Lecture Notes in Computer Science*, Vol. 9042, Springer, pp. 49–65. DOI: 10.1007/978-3-319-18117-2\_4.
5. **Chung, Y.M. & Lee, J.Y. (2001).** A corpus-based approach to comparative evaluation of statistical term association measures. *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 4, pp. 283–296. DOI: 10.1002/1532-2890(2000)9999:9999.
6. **Church, K.W. & Hanks, P. (1990).** Word Association Norms, Mutual Information and Lexicography. *Proceedings of 27th Association for Computational Linguistics (ACL)*, Vol. 16, No. 1, pp. 22–29.
7. **Ebrahimi, S. & Toosi, F.L. (2013).** An analysis of English translation of collocations in Sa'di's Orchard: A comparative study. *Theory and Practice in Language Studies*, Vol. 3, No. 1, pp. 82–87.
8. **Evert, S. (2005).** *The statistics of word cooccurrences: Word pairs and collocations*. Ph.D. thesis, University of Stuttgart.
9. **Evert, S. & Krenn, B. (2001).** Methods for the qualitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pp. 188–195. DOI: 10.3115/1073012.1073037.
10. **Firth, J.R. (1957).** A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, pp. 1–32. Oxford, Philological Society. Reprinted in **Palmer, F.R. (Ed.)** (1968). *Selected Papers of J.R. Firth 1952–1959*, London, Longman.
11. **Gelbukh, A., Sidorov, G., Han, S.-Y., & Hernández-Rubio, E. (2004).** Automatic Enrichment of Very Large Dictionary of Word Combinations on the Basis of Dependency Formalism. *Proceedings of MICAI 2004, Lecture*

Notes in Computer Science, Vol. 2972, pp. 430–437. DOI: 10.1007/978-3-540-24694-7\_44.

12. **Gries, S.T. (2013).** 50-something years of work on collocations: what is or should be next. *International Journal of Corpus Linguistics*, Vol. 18, No. 1, pp. 137–166. DOI: 10.1075/ijcl.18.1.09gri.

13. **Hausmann, F.J. (1979).** Un dictionnaire des collocations est-il possible? *Travaux de Linguistique et de Littérature Strasbourg*, Vol. 17, No. 1, pp. 187–195.

14. **Hausmann, F.J. (2004).** Was sind eigentlich Kollokationen. *Wortverbindungen-mehr oder weniger fest*, pp. 309–334.

15. **Hazelbeck, G. & Saito, H. (2010).** A hybrid approach for functional expression identification in a Japanese reading assistant. *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications*, Beijing, pp. 80–83.

16. **Hoey, M. (2005).** *Lexical Priming: A New Theory of Words and Language*. Psychology Press.

17. **Kim, S.N. & Kan M.-Y. (2009).** Re-examining automatic keyphrase extraction approaches in scientific articles. *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP, ACL and AFNLP*, Suntec, Singapore, pp. 9–16.

18. **Kornai, A. (2008).** *Mathematical Linguistics*. London: Springer-Verlag Limited.

19. **Liddell, D. (1976).** Practical tests of 2x2 contingency tables. *The Statistician*, pp. 295–304.

20. **Michelbacher, L., Evert, S., & Schütze, H. (2007).** Asymmetric association measures. *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2007)*, Bulgaria, Borovets.

21. **Michelbacher, L., Evert, S., & Schütze, H. (2011).** Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory*, Vol. 7, No. 2, pp. 245–276. DOI: 10.1515/cllt.2011.012.

22. **Pakray, P., Neogi, S., Bhaskar, P., Poria, S., Bandyopadhyay, S., & Gelbukh, A. (2011).** A textual entailment system using anaphora resolution. *System Report. Text Analysis Conference Recognizing Textual Entailment Track Notebook*, Vol. 2011.

23. **Pakray, P., Pal, S., Poria, S., Bandyopadhyay, S., & Gelbukh, A. (2010).** JU\_CSE\_TAC: Textual entailment recognition system at TAC RTE-6. *System Report, Text Analysis Conference Recognizing Textual Entailment Track (TAC RTE) Notebook*, Vol. 2010.

24. **Pakray, P., Poria, S., Bandyopadhyay, S., & Gelbukh, A. (2011).** Semantic textual entailment recognition using UNL. *Polibits*, Vol. 43, pp. 23–27.

25. **Pecina, P. (2010).** Lexical association measures and collocation extraction. *Language Resources and Evaluation*, Vol. 44, No. 1–2, pp. 137–158. DOI: 10.1007/s10579-009-9101-4.

26. **Pecina, P. & Schlesinger, P. (2006).** Combining association measures for collocation extraction. *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, pp. 651–658.

27. **Poria, S., Agarwal, B., Gelbukh, A., Hussain, A., & Howard, N. (2014).** Dependency-based semantic parsing for concept-level text analysis. *International Conference on Intelligent Text Processing and Computational Linguistics, Lecture Notes in Computer Science*, Vol. 8403, Springer, pp. 113–127. DOI: 10.1007/978-3-642-54906-9\_10.

28. **Poria, S., Cambria, E., & Gelbukh, A. (2015).** Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. *Proceedings of EMNLP 2015*, pp. 2539–2544.

29. **Poria, S., Cambria, E., Gelbukh, A., Bisio, F., & Hussain, A. (2015).** Sentiment data flow analysis by means of dynamic linguistic patterns. *IEEE Computational Intelligence Magazine*, Vol. 10, No 4, pp. 26–36. DOI: 10.1109/MCI.2015.2471215.

30. **Poria, S., Cambria, E., Howard, N., Huang, G. B., & Hussain, A. (2016).** Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, Vol. 174, pp. 50–59. DOI: 10.1016/j.neucom.2015.01.095.

31. **Poria, S., Cambria, E., Hussain, A., & Huang, G.B. (2015).** Towards an intelligent framework for multimodal affective data analysis. *Neural Networks*, Vol. 63, pp. 104–116. DOI: 10.1016/j.neunet.2014.10.005.

32. **Poria, S., Cambria, E., Ku, L.W., Gui, C., & Gelbukh, A. (2014).** A rule-based approach to aspect extraction from product reviews. *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*, pp. 28–37.

33. **Poria, S., Cambria, E., Winterstein, G., & Huang, G.B. (2014).** Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, Vol. 69, pp. 45–63. DOI: 10.1016/j.knosys.2014.05.005.

34. **Poria, S., Gelbukh, A., Agarwal, B., Cambria, E., & Howard, N. (2013, November).** Common sense knowledge based personality recognition from text. *Mexican International Conference on Artificial Intelligence, MICAI, Lecture Notes in Artificial*

Intelligence, Vol. 8266, Springer, pp. 484–496. DOI: 10.1007/978-3-642-45111-9\_46.

35. **Poria, S., Gelbukh, A., Cambria, E., Hussain, A., & Huang, G.B. (2014).** EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems*, Vol. 69, pp. 108–123. DOI: 10.1016/j.knosys.2014.06.011.

36. **Poria, S., Gelbukh, A., Hussain, A., Howard, N., Das, D., & Bandyopadhyay, S. (2013).** Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, Vol. 28, No. 2, pp. 31–38. DOI: 10.1109/MIS.2013.4.

37. **Rychlý, P. (2008).** A lexicographer-friendly association score. *Proceedings of Recent Advances in Slavonic Natural Language Processing (RASLAN)*, pp. 6–9.

38. **Shin, D. & Nation, P. (2008).** Beyond single words: The most frequent collocations in spoken English. *ELT journal*, Vol. 62, No. 4, pp. 339–348.

39. **Smadja, F.A. (1989).** Lexical co-occurrence: The missing link. *Literary and Linguistic Computing*, Vol. 4, No. 3, pp. 163–168.

40. **Tutin, A. (2008).** For an extended definition of lexical collocations. *Proceedings of the XIII Euralex*

*International Congress*, Barcelona, Spain, pp. 1453–1460.

**Olga Kolesnikova** holds the M.Sc. in Linguistics and the Ph.D. in Computer Science. She is a full-time professor and researcher at the Superior School of Computer Science of the National Polytechnic Institute, Mexico. She is a member of the National System of Researchers of Mexico (SNI 1). Her interests are in computer linguistics and natural language processing, semantic analysis of collocations and other types of restricted lexical co-occurrence, comparative phonetics, and intelligent tutor systems. She authors various publications including a book, a chapter, and articles in international journals. She leads and participates in various research projects on natural language processing, serves as a reviewer in international journals and conferences.

*Article received on 18/01/2016; accepted 25/03/2016.*  
*Corresponding author is Olga Kolesnikova.*