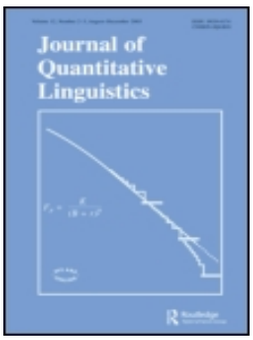

Routledge
Taylor & Francis Group

Journal of Quantitative Linguistics

ISSN: 0929-6174 (Print) 1744-5035 (Online) Journal homepage: http://www.tandfonline.com/loi/njql20

# Measuring Collocation Tendency of Words

Senem Kumova Metin & Bahar Karaoğlan

**To cite this article:** Senem Kumova Metin & Bahar Karaoğlan (2011) Measuring Collocation Tendency of Words, Journal of Quantitative Linguistics, 18:2, 174-187, DOI: 10.1080/09296174.2011.556005

**To link to this article:** https://doi.org/10.1080/09296174.2011.556005

Journal of Quantitative Linguistics
2011, Volume 18, Number 2, pp. 174–187
DOI: 10.1080/09296174.2011.556005

Routledge
Taylor & Francis Group

# Measuring Collocation Tendency of Words*

Senem Kumova Metin[1] and Bahar Karaoğlan[2]
[1]Izmir University of Economics, Balçova-Izmir, Turkey; [2]Ege University, Bornova-Izmir, Turkey

## ABSTRACT

In all natural languages, some words collocate with other words to create multi-worded blocks of meaning – the collocations. Since identification of collocations is vital for information retrieval, language learning, psycholinguistics, authorship determination and translation, collocation extraction is an important issue in natural language processing. In this paper we present a method which is designed to improve current statistical methods that generate ranked lists of collocation candidates.

Due to meaning integrity, any word in a collocation must suggest or at least imply the subsequent words composing the collocation. As a result, we may state that the words in a random text differ in the tendency to facilitate the prediction of the next word. If a word helps the prediction then it tends to collocate, otherwise it does not. In this paper, an attempt has been made to extract collocations by measuring collocation tendency of words and word combinations. The method used is to filter out free word pairs (the words that do not facilitate the prediction of the next word or those in which meaning integrity has not been completed yet) in the lists of candidate pairs.

Collocation tendency method is tested on a base data set extracted by some statistical collocation extraction techniques (frequency of occurrence, point-wise mutual information, the *t*-test, chi-square techniques) and is evaluated by precision and recall measures. We have found that collocation tendency method brings a remarkable improvement on frequency of occurrence and the *t*-test techniques.

*Address correspondence to: Senem Kumova Metin, 35330, Faculty of Engineering and Computer Sciences, Department of Software Engineering, No. 156, Balçova-İzmir, Turkey. Tel: +90232 4888360. E-mail: senem.kumova@ieu.edu.tr. Bahar Karaoğlan, International Computing Institute (ICI), 35100 Bornova-İzmir, Turkey.
Email: bahar.karaoglan@ege.edu.tr.

## 1. INTRODUCTION

A collocation is a recurrent combination of words that co-occur in language more often than by chance to produce natural-sounding speech and writing. Many linguists have given different definitions of collocation, including Firth (1957), Haas (1966), Halliday (1961), Lyons (1966), McIntosh (1966), Van Buren (1967), Leech (1974), Newmark (1978), Benson (1990), Hoey (1991) and Sinclair (1991). Though each definition clarifies particular properties of collocation, there are no known rules that define which words collocate and how a word chooses a particular word or words from millions of different words in language to create a collocation. For example, ''strong'' is a common collocation with ''coffee'' in English, but there is no clear explanation for the preference of this word instead of ''powerful''. In addition, collocations which express the same concept may vary across different languages. In German, ''strong coffee'' is called ''*starker Kaffee*'', although both ''powerful'' and ''strong'' are translation equivalents of the German ''*stark*''. In Turkish, it is ''*sert kahve*'' where the exact translation of the words to English is ''hard coffee''.

Although the definition and the preferences of words in collocation construction seem arbitrary, the properties of collocations have been clearly defined in many previous studies (Manning & Schütze, 2000; Smadja, 1993; Bisht et al., 2006). The easiest and most widely measured property is the recurrence property. As a result, almost all extraction techniques suggest that a collocation must differ from other word combinations in some kind of frequency measure and discriminates between collocations based on frequency of word occurrences (Church & Hanks, 1990; Hindle, 1990; Dunning, 1993; Smadja, 1993; Bisht et al., 2006; etc.).

The other commonly accepted property is the meaning integrity of collocations, which enables collocations to create unit blocks in language. A unit block in natural language is a single word or word combination that has an individual meaning. As a result, if a word combination is a collocation, the composing words may not be treated individually. The idea of meaning integrity in collocations also supports the property of limited compositionality. The compositionality of a natural language expression shows how the meaning of each word changes to create a combination (Manning & Schütze, 2000). If the individual words in the expression lose their own generally accepted meanings, as in idioms, then the expression is regarded as non-compositional. In collocations, words do not completely change their meanings so the meaning of a particular collocation may be predicted to a certain extent. This property in conjunction with the property

of meaning integrity clarifies why a word in a collocation often implies or suggests the rest of the words. The final property, which has been the focus of many previous studies, is the domain and language dependency of collocations. Since collocations are arbitrary, the same concept may be expressed in different ways in different languages and simple word combinations may be used as collocations in different domains.

Due to the property of recurrence, collocations occupy a non-excludable proportion in text and speech. Moreover, due to the property of limited compositionality they have important effects on meaning. As a result, collocation extraction serves for a wide range of natural language processing applications based especially on meaning in text and speech, such as word sense disambiguation, natural language generation and machine translation.

Collocation extraction techniques can be categorized in two main groups: statistical and rule-based methods. Rule-based methods use a large group of rules and require pre-processing steps, such as part-of-speech tagging, to extract collocations. So, they have higher time complexities relative to statistical methods.

Statistical methods (frequency measure, mutual information, hypothesis testing, Smadja's Xtract [1993], the techniques of Kita et al. [1994] and Shimohata et al. [1997]) depend on some kind of frequency measure to ensure the recurrence property and extract frequently used word combinations as collocations. A wide range of commonly used statistical techniques such as frequency measure, mutual information (Church & Hanks, 1990) and hypothesis testing, generate a ranked list of collocation candidates and suggest that the higher the rank, the closer the candidate is to being a collocation. Although list-generating techniques ensure the property of recurrence, we consider that they do not adequately deal with the meaning integrity property. As a result, non-collocated word combinations which occur frequently in the text or speech cannot be discriminated and find their way into the candidate lists.

In this study, we propose a method to filter false collocation candidates by measuring the collocation tendency of words and word combinations in the ranked lists of collocation candidates. The method depends on the idea that if a word combination is a collocation, then any word in the collocation must imply the other words, or at least the subsequent word; due to meaning integrity.

We state that collocation tendency is the measure which shows how much a particular word or word combination helps to predict the

following word, and the tendency is based on the distribution of the subsequent words for the word or combination under consideration in the corpus.

In Section 2, the statistical methods implemented to produce the candidate list are briefly described. In Section 3 we presented the method measuring collocation tendency. In Section 4, experimental set-up, which clarifies the base set and evaluation method, is given. Section 5 involves the implementation results and discussion of our study.

## 2. STATISTICAL COLLOCATION EXTRACTION METHODS

Within the crowded group of statistical collocation extraction methods that generate ranked list of collocation candidates from a given corpus or data set, we give brief descriptions for commonly implemented techniques seen in the literature. In the following subsections; frequency of occurrence, mutual information and some hypothesis tests are defined.

### 2.1 Frequency of Occurrence

The frequency of occurrence method is the simplest and earliest approach to collocation extraction. The words co-occurring more frequently than a given frequency threshold consequently or within a window of words are accepted as collocation candidates. Although some frequent candidates are collocations, others are pairs of function words in the ranked lists of candidates.

### 2.2 Point-wise Mutual Information

Point-wise mutual information is the quantity that measures the mutual dependence of the two words/word combinations (Church & Hanks, 1990). If we write $x$ and $y$ for the first and the second word respectively, point-wise mutual information for them is given by

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x) * P(y)} \tag{1}$$

$P(x)$ is the probability of the word $x$, $P(y)$ is the probability of the word $y$ appearing separately and $P(x,y)$ is the probability of two words $x$ and $y$ coming together in the text.

If the words $x$ and $y$ are independent of each other; the probability of the words coming together must be equal to the product of their own probabilities ($P(x,y) = P(x) \cdot P(y)$). In this case, mutual information will be zero ($I(x,y) = 0$) indicating that the words do not collocate. Therefore, the further the mutual information of a combination moves away from zero, the closer it becomes to being a collocation.

### 2.3 Hypothesis Testing

Hypothesis testing methods test the independence between words in word combinations in order to show that joint occurrences of words in collocations are more than a coincidence. The methods attempt to reject the null hypothesis, which states that words in combination are independent of each other. Among the various different hypothesis tests for evaluating the collocation tendency method, we selected the $t$- and chi-square ($\chi^2$-) tests.

In the $t$-test, the null hypothesis states that the sample is drawn from a normal distribution with mean $\mu$. The test looks at the differences between expected and observed means scaled by variance of the data. As a result, if the observed mean differs from the expected mean significantly, the null hypothesis is rejected. The $t$-statistic is computed as

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}} \tag{2}$$

where $\overline{x}$ is the sample mean (observed mean), $s^2$ is variance, $N$ is the sample size and $\mu$ is the mean of the distribution.

To apply the $t$-test for independence of two words $w1$ and $w2$, we assume that $f(w1)$, $f(w2)$ and $f(w1w2)$ are respective frequencies of $w1$, $w2$ and $w1w2$ and that $N$ is the total number of words in the corpus. Then the following probabilities may be given

$$\begin{aligned} P(w1) &= \frac{f(w1)}{N} \\ P(w2) &= \frac{f(w2)}{N} \\ P(w1w2) &= \frac{f(w1w2)}{N} \end{aligned} \tag{3}$$

Table 1. 2 × 2 table showing observed frequencies for words $w1$ and $w2$.

| | $w1$ | $w1^c$ |
|---|---|---|
| $w2$ | $f(w1w2)$ | $f(w1^c\ w2)$ |
| $w2^c$ | $f(w1w2^c)$ | $f(w1^c\ w2^c)$ |

The null hypothesis is $P(w1w2) = P(w1) \cdot P(w2)$ in the test. If the null hypothesis is true and bigrams[1] of words are generated randomly, then the assignment of 1 to the outcome $w1w2$ and of 0 to any other outcome follows a Bernoulli distribution with mean $\mu = P(w1) \cdot P(w2)$. Using the binomial distribution, the sample mean is $\overline{x} = P(w1w2)$ and sample variance is $s^2 = P(w1w2)(1 - P(w1w2)) \approx P(w1w2)$ since $P(w1w2)$ is small for most word combinations. The $t$-values for all word combinations are calculated and compared with the $t$-value at a predefined level of significance. The null hypothesis is rejected for the combinations that have higher $t$-values than the values in $t$-table.

The $\chi^2$-test technique does not require normally distributed probabilities as in the $t$-test. The test is applied to 2 × 2 tables to compare observed frequencies with expected frequencies in order to examine whether the null hypothesis of independence can be rejected.

Table 1 gives the 2 × 2 observed frequency table of words $w1$ and $w2$ where $f(w1^c\ w2)$ is the number of bigrams with the first word not being $w1$; $f(w1w2^c)$ is the number of bigrams with the second word not being $w2$ and $f(w1^c\ w2^c)$ is the number of bigrams with neither the first word being $w1$ nor the second word being $w2$.

The $\chi^2$ statistic sums differences between observed ($O_{ij}$) and expected values ($E_{ij}$) in all cells of the table and scales the differences by the magnitude of the expectation, as follows

$$x^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{4}$$

where $i$ is the row and $j$ is the column index in the table. The expected frequency of each cell is computed from the totals of rows and columns converted into proportions.

[1]A group of two consecutive words in text or speech.

The $\chi^2$ value is calculated for all word combinations in the corpus and a ranked $\chi^2$ list is generated. The combinations having higher values are accepted as being collocations. Although the application of the $\chi^2$-test is recommended for large probabilities, for which the normality assumption of the *t*-test fails, the $\chi^2$-test is problematic in cases where the numbers in the $2 \times 2$ table are small (Manning & Schütze, 2000).

## 3. METHOD: MEASURING COLLOCATION TENDENCY

The basic idea behind the method we propose here lies in the concept of meaning integrity of words composing the collocation. The meaning integrity in a collocation implies the ease of prediction of the rest of the words once the first word is known. This method improves the performance of the current collocation extraction algorithms by filtering out the candidates that do not show the integrity tendency.

If it is accepted that one or more words of collocation implies the rest, then there must be some kind of measure which describes how much a word may help to predict the remaining words. This measure may be regarded as the level of collocation tendency. We believe that individual words in a collocation differ in the degree to which they tend to integrate with other words. Thus we can use this feature to discriminate true collocations in the candidate lists obtained by common statistical techniques. The method may be defined as follows:

> If a word or word combination does not finalize the meaning it bears, it will tend to collocate with some words. In this condition, to facilitate the integration of meaning and prediction of next word with least effort, the following word will not vary as much as after words with fulfilled meaning. That is to say, collocation tendency of a word or word combination may be defined as the proportion of its frequency to a number of its different subsequent words. As a result, a high tendency for the first word and a low tendency for the candidate compared to a given threshold are required for a collocation candidate of two consecutive words to be assigned as a collocation.

To formulize the method, if $f(w1)$, $n(w1)$, $f(w1w2)$, $n(w1w2)$ represent the frequency of word $w1$, the number of different words following $w1$, frequency of collocation candidate $w1w2$ and number of different words

following $w1w2$ respectively, then the collocation tendency for word $w1$ and for candidate $w1w2$ can be derived from the ratios:

$$T(w1) = \frac{f(w1)}{n(w1)} \quad \text{and} \quad T(w1w2) = \frac{f(w1w2)}{n(w1w2)} \tag{5}$$

$T(w1)$ denotes the level of collocation tendency of $w1$. We can say that the higher the value $T(w1)$ the greater the proneness of $w1$ to collocate. $T(w1w2)$ is the level of collocation tendency of the combination $w1w2$. Hence, the lower the value $T(w1w2)$, the closer the combination $w1w2$ to be a two-word collocation. If $T(w1w2)$ is higher than a given threshold, it denotes that the integrity is not completed and the words in collocation candidate are prone to the creation of $N$-word collocations ($N = 3, 4, 5, \ldots$).

The associative tendency of words is initially discussed by Altmann (1988). He has derived a combinatorial method which measures the significance of association between two words depending on co-occurrence frequencies. Tuzzi et al. (2010) revealed semantic association between non-hapax auto-semantic words in the end-of-year speeches of the Presidents of the Italian Republic using this method. The proposed method in this study can be seen as a variation of Altmann's method which focuses on extracting collocations of two consecutive words.

In this study, we focused on extracting collocations of two consecutive words, but the method is also applicable for $N$-word collocations.

## 4. EXPERIMENTAL SET-UP

### 4.1 The Base Data Set

Collocations are said to be frequently occurring word combinations. Therefore, statistical collocation extraction methods require collocation tagged corpus of great size. Since it is impossible to tag manually all collocations in a large corpus, we prefer to extract a base set from corpora and implement the methods on this set as in some previous studies (Evert & Krenn, 2001; Pearce, 2002). The base set may be constructed from a specific word combination considering as in the study of Evert and Krenn (2001) or the set may be retrieved from a dictionary (Pearce, 2002).

In base data set construction, we intend not only to generate a data set in which all methods may be compared, but also to eliminate all pre-processing steps (e.g. parts-of-speech tagging) used in previous approaches. We retrieved all bigrams in the corpus, excluding those across sentence boundaries, and eliminated word combinations occurring less than five times in the corpus in order to focus on frequently occurring items. Following this, the statistical techniques mentioned in Section 2 were applied to generate a ranked list of bigrams. In the ranked lists, candidates having higher scores were assumed to be collocations. We selected the first (best) 200 candidates in each list to create the base set and tagged the collocations in this set manually. Since the definition of collocation is still a controversial issue, we tagged compound verbs (e.g. take over), frequently used rigid noun phrases, domain specific terms (e.g. strong coffee, heart attack), frequently used word combinations and conjunctions (e.g. ad hoc, strong enough, no way), named entities, personal names, job titles and abbreviations (e.g. New York, general manager, Prof. Dr.) as collocations.

### 4.2 Evaluation Method

Evaluation of extraction techniques generating ranked lists of collocation candidates is performed by recall and precision which are frequently used as performance measures in the field of information retrieval. For collocation extraction, recall may be defined as the fraction of the collocations in the corpus or the base set that is successfully retrieved. Precision is the fraction of true collocations in the retrieved list of collocation candidates. Taking $\varepsilon$ to be collocations extracted from base set and $\delta$ to be the number of true collocations in the base set, recall $r$ and precision $p$ may be defined as

$$r = \frac{|\varepsilon \cap \delta|}{|\delta|} \qquad p = \frac{|\varepsilon \cap \delta|}{|\varepsilon|} \tag{6}$$

While presenting the precision and recall values, instead of computing the measures for only a single proportion of candidate list (for example just for the whole set or just for the first $N$ candidates), we computed recall and precision for $N$ highest ranked candidates where $N$ may vary from 1 to the total number of candidates ($N = 1, 2, 3, \ldots$, base data set) as in the study of Evert and Krenn (2001). In this manner, we prevent

misleading conclusions being drawn from a single value of $N$. Since we have plotted graphs of precision and recall for the whole base set, recall values maximize to 1.

## 5. RESULTS AND DISCUSSION

We utilized a part of Leipzig Corpora Collection, which is compiled for scientific use by Leipzig University Department of Natural Language Processing (Quasthoff et al., 2006). Collocation extraction techniques were applied to 100,000 sentences in English; the utilized part has 630 distinct and 216,206 bigrams with a frequency of greater than five.

The base set constructed from the top 200 best candidates of four statistical methods involves 440 distinct candidates; 48.86% of the set are true collocations. The proportion gives us the baseline for precision graphics.

Figure 1 presents the precision curves before the application of collocation tendency method. The horizontal axis presents the percentage of base set completed (the value $N$ in percentages) and the vertical axis is the precision value. It is noteworthy in Figure 1 that chi-square and point-wise mutual information methods give consistently higher precision values compared to frequency of occurrence and the $t$-test methods. In addition, both frequency of occurrence and the $t$-test methods are lower than even the baseline value (0.4886).

Figure 2 gives the initial recall graphs for the methods. The horizontal axis presents the percentage of base set completed (the value $N$ in percentages) and vertical axis is the recall value. Similar to precision

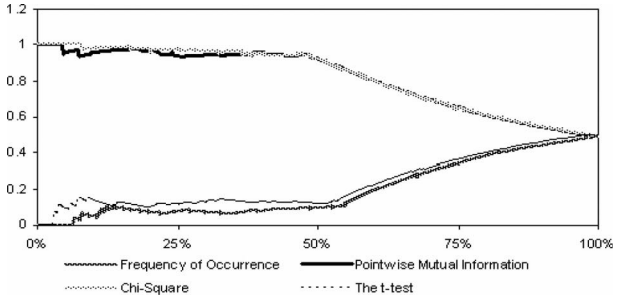


Fig. 1. Precision graph of base set before the application of collocation tendency method.

results, chi-square and point-wise mutual information methods generate higher scores for true collocations and extract them earlier than the other two methods.

In the filtering step, collocation tendency method generated a new base set of 147 collocation candidates by eliminating some of the candidates in the initial set. The collocation tendency threshold is determined by the experiments in which we increment it beginning from zero till we reach both the highest precision and recall for the whole initial base set. It is found to be $(0.7)^{-1}$ after the experiments. The overall precision (baseline) was measured as 0.8776 for the filtered base set. Figures 3 and 4 depict the precision and recall graphs for filtered set respectively. In the figures, the vertical axis presents the precision/recall values; horizontal axis is the completed portion of filtered base set.

The pre- and post-collocation tendency filtering results show a considerable improvement in frequency of occurrence and *t*-test methods; both in precision and recall measures, as a result of elimination of

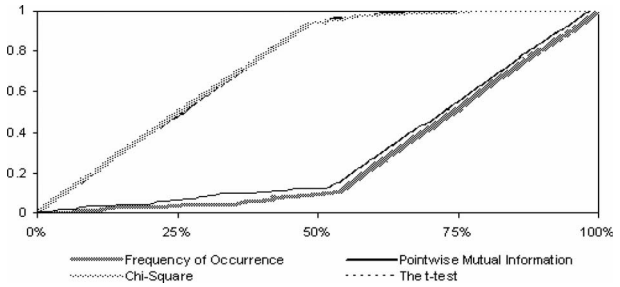


Fig. 2. Recall graph of base set before the application of collocation tendency method.

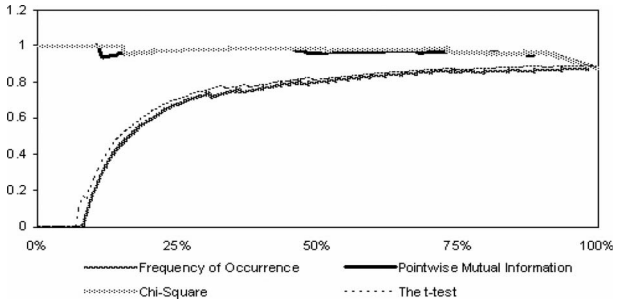


Fig. 3. Precision graph of base set after the application of collocation tendency method.

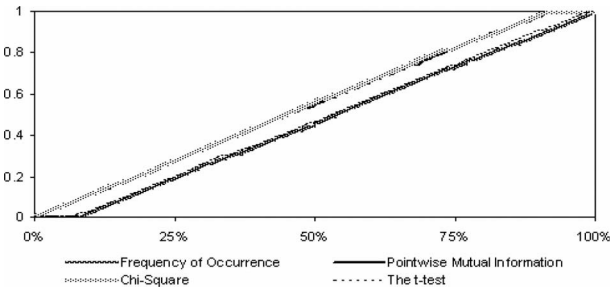


Fig. 4. Recall graph of base set after the application of collocation tendency method.

candidates by the proposed method. In addition, it is seen that our approach considerably reduces the difference between the methods.

One other important result pointed out by the figures is that although precision curves for point-wise mutual information and chi-square tests seem to be better after filtering, it is not possible to state that the method is completely successful. If the initial chi-square and point-wise mutual information precision values of the top 147 candidates is compared with the filtered values of a base set of 147 candidates, we do not see a significant improvement.

As a consequence, it can be stated that the collocation tendency method enables the results of collocation extraction methods to converge by filtering the candidates depending on the concept of meaning integrity. In particular, it improves the effectiveness of two methods to the level of other competing methods in the study: the *t*-test method, which assumes a normally distributed data, as opposed to the chi-square test, and the frequency of occurrence method, which is accepted as the easiest way to extract collocations. Since time complexity is an important concept in natural language processing applications, we believe that the improvement of frequency of occurrence method is an especially valuable contribution. Moreover, the proposed method is able to extract not only two-word collocations but also multi-word collocations if applied recursively.

## REFERENCES

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography*, *3*, 23–34.

Bisht, R. K., Dhami H. S., & Neeraj Tiwari, N. (2006). An evaluation of different statistical techniques of collocation extraction using a probability measure to word combinations. *Journal of Quantitative Linguistics*, *13*, 161–175.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

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*, Toulouse, France (pp. 188–195).

Firth, J. R. (1957). Modes of meaning. *Papers in Linguistics 1934–51*. Oxford: Oxford University Press.

Haas, W. (1966). Linguistic relevance. In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (Eds.), *In Memory of J. R. Firth* (pp. 116–148), London: Longman.

Halliday, M. A. K. (1961). Categories of the theory of grammar. *Word*, *17*(3), 241–292.

Hindle, D. (1990). *Noun Classification from Predicate-Argument Structures*. Annual Meeting of the Association for Computational Linguistics (ACL 1990), Pittsburgh, Pennsylvania, USA (pp. 268–275).

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Justeson, J. S., & Katz, S. M. (1995). Principled disambiguation: Discriminating adjective senses with modified nouns. *Computational Linguistics*, *21*(1), 1–28.

Kita, K., Kato, K., Omoto, T., & Yano, Y. (1994). A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. *Journal of Natural Language Processing*, *1*, 21–33.

Leech, G. (1974). *Semantics*. London: Penguin.

Lyons, J. (1966). Firth's theory of meaning. In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (Eds.), *In Memory of J. R. Firth* (pp. 228–302), London: Longman.

Manning, C. D., & Schütze, H. (2000). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT Press.

Mc Intosh, A., & Halliday, M. A. K. (1966). *Patterns of Languages: Papers in General, Descriptive and Applied Linguistics*. London: Longman.

Newmark, P. (1978). *Approaches to Translation*. Oxford: Pergamon Press.

Pearce, D. (2002). A comparative evaluation of collocation extraction techniques. *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (LREC 2002), Las Palmas, Spain.

Quasthoff, U., Richter, M., & Biemann, C. (2006). Corpus portal for search in monolingual corpora. *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006), Genoa (pp. 1799–1802).

Shimohata, S., Sugio, T., & Nagata, J. (1997). Retrieving collocations by co-occurrences and word order constraints. *The 8th Conference on European Chapter of the Association for Computational Linguistics*, Madrid, Spain (pp. 476–481).

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Smadja, F. A. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, *19*(1), 143–177.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative Analysis of Italian Texts*. Lüdenscheid: RAM-Verlag.

Van Buren, P. (1967). Preliminary aspects of mechanisation in lexis. *Cahiers de Lexicology*, *11*, 89–112; *12*, 71–84.