Dictionary of Lithuanian Phrases

Rūta Marcinkevičienė
Vytautas Magnus University
Donelaičio 58
KAUNAS
LITHUANIA
ruta@hmf.vdu.lt

Abstract
The paper discusses problems related to the compilation of the first Lithuanian dictionary of collocations. The problems encountered are both theoretical and empirical in nature. Theoretically, it is important to differentiate between collocations consisting of nodes and their collocates on one hand and collocational strings, i.e. clear-cut real-text chains of words, on the other. The latter notion is applied as the basis of the method for the extraction of collocations from the corpus of the Lithuanian language. Empirically, the statistical output, i.e. collocations extracted from the corpus of 100 million running words, is described from the point of view of their grammatical form, lexical autonomy and boundaries, and the steps taken to transform the list of statistical collocational strings into the list of linguistically acceptable collocations or phrases are outlined.

1 Two Approaches to Collocation

It was necessary to compile the first dictionary of collocations from the corpus of 100 million running words of Lithuanian texts in order to re-evaluate different approaches to collocation and the methods of extraction. A lexicographic approach to collocations and a modified method for their extraction when applied to the Lithuanian corpus resulted in an output of the so-called statistical collocational strings, which have to be manually processed before they are included in the dictionary of collocations. The paper therefore overviews the approach and the method of extraction, concentrating on the transformation of statistical output into the dictionary of collocations.

Collocation is a fuzzy term embracing a great variety of notions. The definition of collocation differs according to researcher and standpoint. It also depends on the methods of extraction that provide researchers with lists of frequently co-occurring lexical items. There are two different perspectives on the notion of collocations from the point of view of their form and structure. One group of authors (Firth, 1957; Sinclair, 1991; Stubbs, 2001, among others) prefers contextual or statistical definition of collocation. Contextual definition could be generalised as follows: one item collocates with another if it appears somewhere near it in a given text (Partington, 1998: 16). For a statistical definition see Stubbs (2001: 29): "Usually it is frequent co-occurrences which are of interest, and corpus linguistics is based on the assumption that events which are frequent are significant. My definition is therefore a statistical one: 'collocation' is a frequent co-occurrence.". The assumption underlying collocation is based on the structural notion of a collocation, e.g.: a collocation consists of a node word and its collocates, so the search for a collocation starts with the node word. Therefore, most collocations are usually constructed as binary items consisting of a node and its collocates found within a previously selected span.

This view of collocation as a relationship between two or more words is reflected in the way collocations are presented. The most convenient notation for presenting information on collocations is a list of collocates of a node word. The list contains notional words, function words being omitted: *Caring <seeks, loving, honest, sincere, etc.>*. For full phrases one has to resort to concordances of the node word or reconstruct coherent word combinations using native speaker informants. One example with the node word and its collocate for the above list could be a phrase from the lonely hearts ad: *Seeks sincere, caring single lady.* (cf. Stubbs, 2001: 19). This approach employs the term for the description of lexical relation rather than for real word combinations: "Collocation is a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text, e.g. *provide* occurs with *help, assistance, money, food, shelter, information.*" (Stubbs, 2001: 24).

Highlighting lexical relationship, the statistical notion of a collocation as it were disrupts the real language specimen, i.e. extended lexical units or strings of words, e.g. *provide help* and *provide shelter*. Moreover, it differentiates structural components of a collocation hierarchically, distinguishing a node word and its collocates, which presumes a list of node words previous to the extraction of collocations. Furthermore, it does not allow one to detect multi-word collocations or to define their boundaries.

Consequently, without a clear idea of the boundaries of collocations it is impossible to determine definitively which part of a corpus or its subcorpus consists of collocations and, on a larger scale, which part of a particular variety of a language is formed on the idiom principle (Sinclair, 1991: 109-121). Concerning the boundaries of collocations Sinclair states that "The boundaries between stretches constructed on different principles will not normally be clear-cut...nevertheless it is possible to measure statistically the length of collocations and to set its boundaries (Sinclair, 1991: 113).

Another group of authors (Kjellmer, 1982; Williams, 1998, among others) pursue a lexicographic approach and include grammatical well-formedness in the list of criteria of collocations. Collocations for them are not purely statistical. "If frequency alone were to be our guide in extracting collocational material from the Corpus, it is clear that the material would be of a very heterogeneous nature. Moreover, it is obvious that if the inventory is to have the form of a dictionary, the *although he, but too, hall to* type of combinations is of very limited value;" (Kjellmer, 1982: 25). The association of frequency and grammatical structure proved to be a means of selecting those combinations that qualify for inclusion in the dictionary of collocations.

The statistical approach, however, does not include grammatical acceptability among the criteria of collocation. On the contrary, a radical approach concerning the form of collocation is presented by Sinclair, who rejects the necessity for collocation to be interpreted from the point of view of its grammatical structure: "Just as it is misleading and unrevealing to subject *of course* to grammatical analysis, it is also unhelpful to attempt to analyze grammatically any portion of text which appears to be constructed on the idiom principle." Constraints other than grammatical ones are placed on those strings of words that constitute single choices and can be called preconstructed or semi-preconstructed phrases. "The boundaries between stretches constructed on different principles will not normally be

clear-cut, and not all the stretches carry as much evidence as *of course* does to suggest that it is not constructed by the normal rules of grammar." (Sinclair ,1991: 113).

An additional argument against the specific grammatical form that a collocation should adhere to is the fact that collocations are formed by semantic units irrespective of grammatical category: *argue heatedly*, *heated argument*, *in the heat of argument* (Stubbs, 2001: 30). The only grammatical categories relevant for collocation from this point of view are morphological forms of a node word which are reported to demonstrate idiosyncratic collocability.

One way to solve the problem of a fuzzy notion of collocation and to differentiate between the two approaches would be to use different terminology, i.e. to reserve the traditional term *collocation* to the relationship between two or more words, which are presented in the form of a list, and to use the term *statistical collocational strings* for authentic chains of words extracted from the corpus. The latter term would stress the cohesion of the raw output after the application of a particular method of extraction. The manually processed definitive version would list the *most frequent phrases*.

A compiler of a dictionary of collocations has to choose between the two approaches, since the attitude towards collocation predetermines the method of extraction and the method of presentation. From the perspective of the Lithuanian language the lexicographic approach is more acceptable. It provides a lexicographer with authentic strings of words obtained by applying statistical tools. These strings contain collocating grammatical forms presented in their natural word order, thus not isolated lemmas, which are of paramount importance for the highly inflected Lithuanian language. Collocational strings can be sorted with their grammatical autonomy in mind but they do not have to be reconstructed from a mere list of nodes and their collocates.

Finally, this approach allows us to avoid making a pre-selected list of node words and to process the entire corpus from the first to the final word. It presents, therefore, a full-text approach to language and utilises the entire corpus, i.e. every sentence it contains, not merely concordances derived from the corpus on the basis of a previously compiled list of node words. Thus calculations of collocability are applied to the continuous chain of words. Consequently this approach allows us to determine the amount of text that is formed on the idiom principle (Sinclair,1991: 109-121). The choice of the lexicographic approach as opposed to the statistical one informs the choice of a particular method for the extraction of collocations.

2 Gravity Counts as the Method of Extraction

Collocational strings were extracted from the corpus of Lithuanian language with the help of a statistical method called Gravity Counts. It adopts a linear approach of consecutive counts of words in a text, and of all the texts in a corpus, based as it is on the combinability counts of each pair of words in the corpus irrespective of their hierarchical status, i.e. there is no *a priori* list of node words for which collocates are obtained from the corpus. Each word in the corpus is processed as the node word; its gravity by reference to the pairing word and the next two words in the span of three words is calculated using the formula below (for more about the method see Daudaravičius, Marcinkevičienė, forthcoming):

$$G(x,y) = \log\left(\frac{f(x,y)\cdot n(x)}{f(x)}\right) + \log\left(\frac{f(x,y)\cdot n'(y)}{f(y)}\right)$$

Gravity Counts are based on an evaluation of the combinability of two words in a text that takes into account a variety of frequency features, such as individual frequencies of words, the frequency of a pair of words and the number of different words in the selected span. Gravity Counts highlight habitual co-occurrence of two words in a text within the chosen span, in our case the span of three words. If the first word *x* is used more habitually than expected in front of the second word *y*, and the second word *y* is used more habitually than expected after the first word *x*, then *x* and *y* form a minimal collocational string.

Gravity Counts are also based on word order, so that for each first word *x* in a pair the frequency of the following three words is taken into consideration, while for each second word *y* of a pair the frequency of the three preceding words is computed. Therefore *n(x)* is the number of different words to the right of *x* and *n'(x)* words to the left of *y*; *f(x)* and *f(y)* is the frequency of *x* and *y* in the corpus.

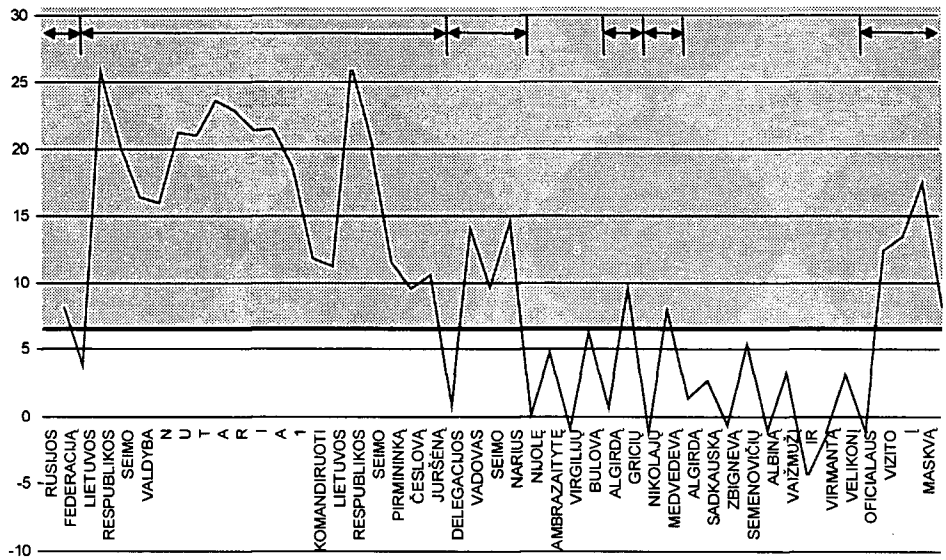

Figure 1: A sentence from the corpus presented as a curve of combinability

While the corpus is processed using Gravity Counts the span of three words is moved along the lines of the corpus with the aim to detect statistically reliable collocational strings of words. Extraction of strings of various length allows us not only to detect the relationship of collocation between words of the corpus, but, using a text as the basis for the extraction of collocations with the help of the Gravity Counts, to detect the statistical boundaries of each collocation. A collocational string is statistically defined as a segment of text where the combinability of constituent adjacent word pairs is above the arbitrarily chosen point of collocability. The lower combinability of word pairs preceding and following the segment (as well as the beginning and the end of a text or a corpus) marks the boundaries of a collocational string.

The core of the definition of a collocational string in our case is the diversity count of its lexical surroundings. The coefficient of diversity shows the relationship between the word in question and diversity of the words surrounding it. The higher the frequency of the pair of words over the standard diversity, the higher the value of their combinability and vice versa.

Figure 1 exemplifies the method applied to the corpus of Lithuanian language seen as a changing curve of lexical combinability. The fragment of one sentence shows peaks appearing above the arbitrary boundary of 5.5, taken as collocational strings consisting of a different number of words. Locating the boundaries of collocational strings gives us a unique opportunity to analyse their representation in the corpus, in other words, to establish what part of the corpus is made up of collocations. Furthermore, it becomes possible to measure the average length of the strings, and the relationship between the length of a collocational string and its frequency.

3 The Output

com Application of Gravity Counts for the corpus of Lithuanian language resulted in processing of 110 935 pairs in the corpus of 100 million running words (1,7 million different word forms). Some pairs of words were joined into multi-word collocational strings, thus the overall list of collocational strings consists of 19,878,281 items. The list of different items is of 10,147,250 items. All the collocations cover 68.1 percent of the corpus. This number is comparable to Altenberg's (1991) results. He showed that about 70 per cent of the words of

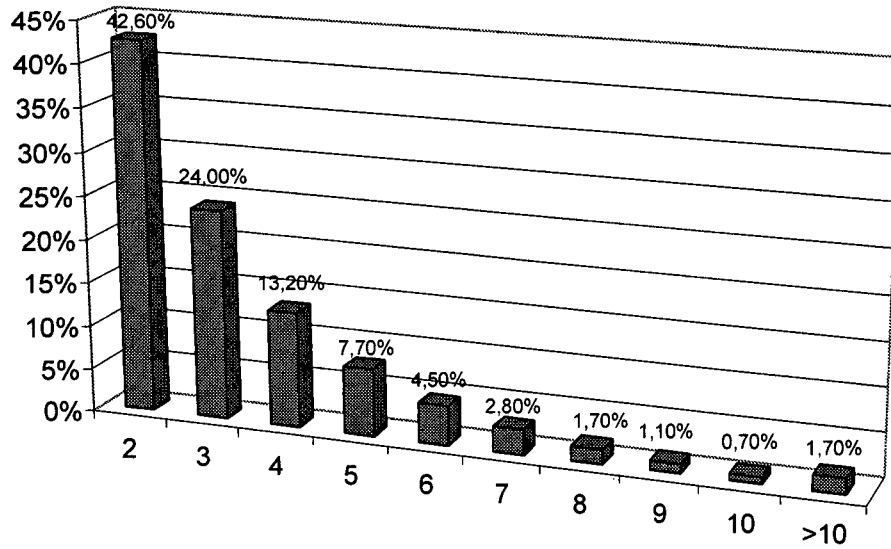

Figure 2: Distribution of collocational strings by their length (in number of words)

running text in the *London-Lund Corpus* (the size of which is half a million words) are used in recurrent word binations.

The output of the calculations is the list of collocational strings of varying length. The general tendency for the length of collocations is the same as for the frequency of words, i.e. the longer collocations are less frequent than the shorter ones. The majority of collocations (8,462,626 items which form 42 per cent of all the list) are made up of two words. The list of three-word collocations is twice as short (4,760,991 items, 24 per cent of the list), the same can be said of the four word collocations (2,629,953 items, 13 per cent of the list) and the five word collocations (1,532,370 items, 8 per cent of the list). The decrease in number for the longer collocations is somewhat less (see Figure 2). A typical long collocation is taken from governmental decrees and consists of 34 words.

From the point of view of coverage, the general tendency is for less frequent but lengthier collocations to comprise the major part of the corpus (see Figure 3). The collocational strings used only once comprise 39.1 per cent of the corpus. Frequent collocational strings, occurring in the corpus more than ten times, form a relatively small part, i.e. 14.5 per cent. This means that only one eighth of our language formed on the idiom

principle consists of frequently used collocations. The remaining part of the idiomatic language consists of relatively rare collocational strings.

Statistical criteria are sufficient to identify collocational strings. Phrases or well-formed collocations, however, demonstrate other features typical of collocations. A tighter definition describes collocation not only as habitual, but also as grammatical, meaningful, arbitrary, lexically transparent, language-specific, or consisting of at least two notional words, to mention the most important features of a well-formed collocation (Kennedy, 1998; Partington, 1998; Hunston and Francis, 2000, among others). In our case the first three features are the most important. The statistical method applied guarantees that the segments

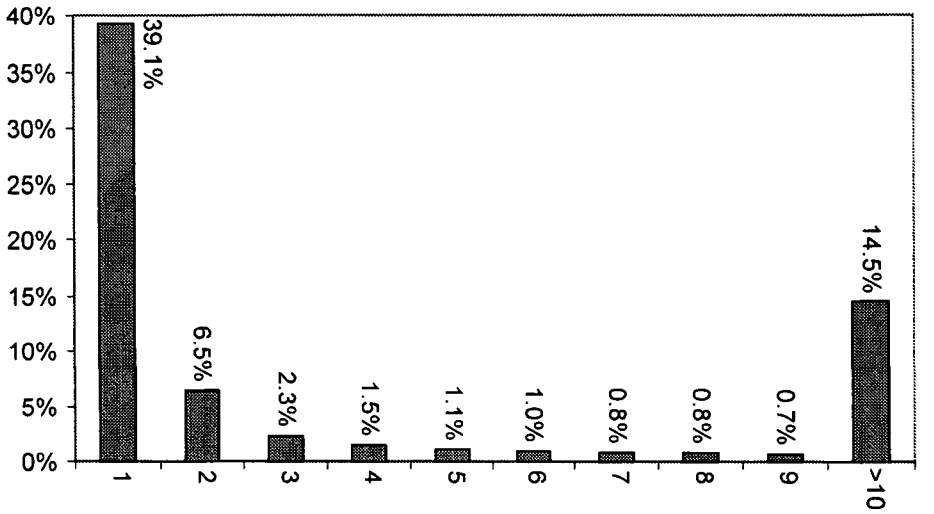

Figure 3: Distribution of collocational strings by their frequency

of text are habitual, so that one needs only analyse the data obtained from the grammatical and semantic point of view, i.e. discover whether statistical boundaries of collocation coincide with the boundaries of a linguistic unit.

The method of Gravity Counts and the detection of collocation boundaries helps to identify segments of texts as statistically significant chains of words. These chains can be said to be always natural since they present authentic fragments of a text. Nevertheless, statistical collocational strings differ from the point of view of their grammatical and lexical autonomy, which is the most relevant feature in our analysis. Certain collocational strings are self-sufficient and can be regarded as autonomous and grammatically well-formed phrases, e.g. *visų pirma* (first of all), *praėjusią savaitę* (last week), *dar kartą* (once more), *kitaip tariant* (in other words), *kitą dieną* (next day), *ilgą laiką* (for a long time), *šį kartą* (this time), etc. These phrases are used in the corpus in various morphological forms, e.g. *informacijos šaltiniai, informacijos šaltiniais, informacijos šaltinio, informacijos šaltinis, informacijos šaltiniu, informacijos šaltiniams, informacijos šaltinių* (the source of

information used in different case forms). In spite of the formal variety, collocational strings used in their non-lemma forms can be treated as phrases.

Other types of autonomous phrases include the so called morphological clusters: a) complex particles that are comprised of two separate, semantically indivisible, components: *vos tik* (hardly), *ko tik* (just,) *kad ir* (even); or retain their own meaning but are predominantly used together: *lyg tai* (it seems), *vien tik* (just only); b) compound tense and voice forms: *turėtų būti* (should be), *gali būti* (can be), *buvo sumažinta* (was decreased); c) compound pronouns as *kai kas*, *kas nors* (something, somebody), *bet kas* (anything, anybody), pronominal groups: *toks pat* (the same), *beveik visas* (almost all), *visiškai kitas* (quite another); d) idiomatic prepositional phrases: *be abejo* (no doubt), *iš karto* (at once), *iš tiesų* (to tell the truth), *be galo* (extremely) or non-idiomatic but fossilised prepositional phrases *iš pradžių* (from the begining), *į priekį* (forward), *iš gailesčio* (out of pity), *iš viso* (in sum); e) a great variety of other phrases common in the media and other genres: *vienaip ar kitaip* (one way or another), *ilgą laiką* (for a long time), *norom nenorom* (willy-nilly), etc. The first three types are presented in the grammar of the Lithuanian language as morphological units (Ambrazas, 1997: 396).

All the above-mentioned types of phrases are meaningful, grammatical and clearcut. Other kinds of collocational strings are somewhat deficient. Some of them lack notional words, since they consist of frequent clusters of pronouns and conjunctions: *tarp jų* (between them), *ir jos* (and hers), *su juo* (with him); some are typical parts of sentences lacking continuation, falling out of the range of a collocational segment of a text: *manoma kad* (it is supposed that), *sakė kad* (s/he said that), *priklausomai nuo* (depending on), *aš noriu* (I want), *duomenys apie* (data about), *dar labiau* (even more), etc. Parts of these non-autonomous strings are predictable for a native speaker on the basis of their morphological forms, therefore a full linguistic unit can be generated using either a corpus or one's intuition or both, e.g. *[imti, ėmė, paima] iniciatyvą į savo rankas* ([take, took, takes] the initiative into one's hands), *[sirgti, sirgo, serga] inkstų ligomis* ([to be, is was ill] with kidney deseases).

A distinctive feature of the collocational strings derived with the help of gravity counts is the encapsulation of collocations. Due to the fact that identical strings of various length are detected and counted in the long chain of all the texts in the corpus, parts of the collocations coincide, e.g.:

dalinio pakeitimo (of partial change),

dalinio pakeitimo ir papildymo (of partial change and amendment),

dalinio pakeitimo ir papildymo projektas (the project of partial change and replenishment).

It is possible to construct a full phrase from its fragments by using the alphabetical list. Usually a longer phrase occurs less frequently than the shorter ones that it encapsulates. In the example above, the shortest string is used more often than the longer one, which in turn is more frequent than the longest collocational string. Variation in frequency explains why encapsulated collocations appear in the list separately.

In order to differentiate between autonomous and deficient collocational strings obtained from the corpus using the method of Gravity Counts, as well as to define their ratio, a more detailed analysis is necessarily preceded by a description of criteria for identification

of autonomous collocations. For the lexicographic purpose it suffices to say that a fairly high percentage, i.e. 82 % of collocational strings are autonomous and clear-cut phrases.

4. Transformation of collocational strings into phrases

The manual processing of the raw output, i.e. transformation of statistical collocational strings into well-formed phrases, consists of several steps and procedures. The first step is to delete all rare strings, irrespective of their length (1 to 3 occurrences) and some more frequent strings depending on their length: two-word strings up to 19 occurrences, three word strings up to 9 occurrences, four word strings up to 8 occurrences, five word strings up to 4 occurrences. This arbitrary decision was based on the considerable amount of noise in these particular word groups.

The remaining list of 73,188 collocational strings of different length was processed applying three different procedures: lexically well-formed and grammatically autonomous collocational strings were included without changes. Some strings were deleted (anomalous, insufficient, e.g. parts of the string belonging to a different clause, or strings containing proper names, numbers, misprints, consisting exclusively of a noun plus conjunction, a pronoun or one of the forms of the verb to be) while some were changed. The changes include: a) shortening of grammatically irrelevant parts of long collocations, b) addition of missing words from concordances to deficient strings, mostly two or three word combinations consisting of nouns and prepositions, e.g. *legenda apie* (*kilmę*), or any other parts of speech, e.g. (*peržengė*) *padorumo ribas* c) junction of embedded collocations, e.g. *asmenims, kurie* (occurring 73 times) is embedded into a weaker collocation *asmenims, kurie šią žemę išnuomoja* (occurring 5 times), forming one collocation consisting of components with different respective frequencies, e.g. *asmenims, kurie* 73 *šią žemę išnuomoja* 5. Collocational strings made up of identical common nouns, but differing proper nouns, are joined, leaving the common nouns and the most frequent proper nouns, e.g. *dėl posėdžio vedimo tvarkos kalbėjo seimo narys (A. Baležentis; A Endriukaitis; A Kubilius).*

The manual processing of collocational strings is not yet completed, therefore it is difficult to give definitive numbers and to determine how many statistical collocations are included in the dictionary unchanged, how many of them are deleted and how many slightly transformed. The first stage of transformation, i.e. the deletion of deficient strings, left the compilers of the dictionary with 60,040 items. Since changes in the collocational strings do not affect the length of the list, it is only the junction of similar or embedded collocational strings that can shorten it. To sum up the overall procedure, the initial list of 20 million collocational strings was transformed into 10 million different collocational strings. After deletion of noisy word groups circa 73,000 collocational strings remained for human inspection. It can be predicted that the final list of phrases should contain circa 50,000 lexical items.

5 Concluding remarks

The approach to collocation and the specific method of extraction described here was applied to the corpus of the Lithuanian language, with several aims in mind. First, to apply a new method for the extraction of collocational strings, second, to detect the part of the corpus that is formed by idiomatic language, and finally, to compile an electronic dictionary

of frequent phrases. The completed dictionary may be used not only for the practical purposes of lexicography, statistics, NLP, language learning and as a reference tool on usage but also for further research. Possible areas of research could include comparison of the output of different statistical tools, checking the list against native speakers' intuition for the inclusion of frequently used phrases, to mention a few.

Acknowledgments

I am grateful to Patrick Corness, visiting research fellow at the Centre for Translation Studies, University of Leeds, for his careful reviewing of this text, and his helpful linguistic advice.

References

Altenberg, B. 1991. 'Amplifier Collocations in Spoken English' in S. Johansson and A.B. Strenström (eds). *English Computer Corpora.* Berlin: Mouton de Gruyter.

Ambrazas, V. 1997. *Lithuanian Grammar.* Vilnius: baltos lankos.

Firth, J. 1957. *Papers in Linguistics.* London: Oxford University Press.

Daudaravičius, V., Marcinkevičienė R. (forthcoming).'Gravity Counts for the Boundaries of Collocations' in *International Journal of Corpus Linguistics.*

Hunston, S. and Francis, G. 2000. *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics.* London/New York: Longman.

Kjellmer, G. 1982. 'Some Problems to the Study of Collocations in the Brown Corpus' in Stig Johansson (ed.) *Computer Corpora in English Language Research.* Bergen: Norwegian Computing Centre for Humanities.

Partington, A. 1998. *Patterns and Meanings. Using Corpora for English Language Research and Teaching.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sinclair, J.McH. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Stubbs, M. 2001. *Words and Phrases. Corpus Studies of Lexical Semantics.* Oxford: Blackwell Publishers.

Williams, G. 1998. 'Lexis in a Corpus of Plant Biology Research Articles' in *International Journal of Corpus Linguistics,* vol. 3, No 1.