CORPUS

Corpus

11 | 2012
La cooccurrence, du fait statistique au fait textuel

Les collocations comme indice pour distinguer les genres textuels

Stefania Spina et Elena Tanganelli

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/2219
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 1 janvier 2012
ISSN : 1638-9808

Référence électronique
Stefania Spina et Elena Tanganelli, « Les collocations comme indice pour distinguer les genres textuels », *Corpus* [En ligne], 11 | 2012, mis en ligne le 18 juin 2013, consulté le 03 mai 2019. URL : http://journals.openedition.org/corpus/2219

© Tous droits réservés

Collocations as an index for distinguishing text genres

Stefania SPINA, Elena TANGANELLI
Department of Language Sciences,
University for Foreigners Perugia,
Piazza Fortebraccio 4, 06122 Perugia, Italy

1. Aims and background

The distinction of different registers by computational means has been an ongoing goal of corpus linguistics research since Biber (1988), who used 67 linguistic features operating at the word level in order to study variation in speech and writing. The same goal, using different computational and statistical methodologies, is shared by natural language processing (for example Karlgren & Cutting 1994; Kessler, Nunberg & Schütze 1997, Peng *et al.* 2003).

Biber's (1988) study and its multidimensional approach have become a standard in corpus linguistics, in all its versions and extensions (Conrad & Biber 2001, Lee 2004).

This paper aims to incorporate collocations as an index for distinguishing text genres: our main hypothesis is that collocations, as well as other linguistic features, are potentially suitable to identify genres. Thus, this is mostly an exploratory study, aimed at verifying this hypothesis and at taking a deeper look into register variation across different genres in Italian with computational and statistical methods; this approach is largely underused in studies on Italian, mainly due to the scarcity of comprehensive collections of data that are representative of multiple textual genres.

Furthermore, in a broader perspective, this study might make significant contributions to other fields, such as automatic genre identification (Santini 2004) or measures of text cohesion (Louwerse *et al.* 2004) or text readability, where the detection

Corpus n°11, *La cooccurrence : du fait statistique au fait textuel*
(2012), 73-89

of collocations as marker for genres can increase the accuracy of computational tools devoted to these tasks.

From a theoretical point of view, an assumption shared by numerous fields of research is that language has a clearly phraseological nature: collocations tap into both the paradigmatic and syntagmatic features of texts; they occupy a place at the intersection between lexis and syntax, and their use entails the co-selection of both levels, which consequently cannot be analysed separately. This view is common to different approaches and theoretical frameworks such as Cognitive Grammar (Langacker 1991), Construction Grammar (Goldberg 1995), and Corpus Linguistics (Sinclair 1991).

Collocations, in addition, are one of the elements that contribute to text cohesion (to lexical cohesion in particular: Halliday & Hasan 1976); the way they achieve this task varies across different textual genres, as shown in Laybutt (2009).

We adopt here Biber's definition of genre: "Genre categories are determined on the basis of external criteria relating to the speaker's purpose and topic; they are assigned on the basis of use rather than on the basis of form" (Biber 1988: 170).

By collocation, we mean "the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance" (Gries 2008: 3). This definition emphasizes one of the differences between ours and Biber's approach: he mainly used linguistic features operating at the word level, while we use the co-occurrences of two or more words which function as one semantic unit as a means to identify text genres.

Other corpus studies proceed in the same direction: Biber *et al.* (2004) use lexical bundles to compare conversation and academic prose; Crossley & Louwerse (2007) use frequency of bigrams to classify spoken and written registers; Gries & Mukherjee (2010) and Gries (2009) use lexical gravity of n-grams to analyse differences in Asian English and in different registers of English; Gledhill (1995 and 2000), Spina (2010c) and Simpson-Vlach & Ellis (2010) use collocations in academic

language analysis; Laybutt (2009) studies collocations as a sign of textual cohesion in sports reports.

All these studies are based on the computation of lexical bundles, n-grams or collocations as identifiers of registers or varieties of language, and prove the effectiveness of n-grams as a tool for discriminating linguistic variation (Biber 2009).

In an attempt at a more in-depth investigation of the use of collocations to distinguish among text genres, the following are the two research questions which this study aims to answer:

- are collocations a suitable index for distinguishing text genres?
- are there statistical measures that are particularly appropriate for this task?

The article is organized as follows: in section 2 the methodology used is described (in particular, section 2.1 illustrates the criteria used for extracting the collocations and section 2.2 introduces lexical gravity, the statistical measure on which this study is based; section 2.3 describes the corpus and the different text genres from which the collocations have been extracted, and provides some data on the number of final collocations obtained); in section 3, the results of the application of raw frequency and lexical gravity to the collocations studied are discussed, as is the statistical test (cluster analysis) applied to these two measures to explain the various distributions throughout the different text genres. Section 4 presents some final considerations.

2. Methodology

2.1 Extraction of collocations

From a methodological point of view, this study differs from previous ones in two main aspects.

First of all, collocations are extracted as sequences of parts-of-speech (on the effectiveness of this approach: Pazos Bretana & Pamies Bertran 2008; Evert 2008). According to the methodology discussed in Spina (2010a and 2010b), the candidate collocations are extracted based on the most frequent combinations of grammatical categories in Italian, which give as a result a sequence of linguistic elements "which functions as one semantic unit" (Gries 2008: 3). We have selected 4 of the most productive POS sequences in Italian: verb-noun (VN),

noun-noun (NN), noun-preposition-noun (NpreN), noun-adjective (NADJ). The following are examples of each sequence:

> VN: fare la doccia [take a shower], avere bisogno [need/have need] (avere molto bisogno [have great need], avere veramente un gran bisogno [really have great need]);
>
> NN: fine settimane [weekend], sito web [website];
>
> NpreN: punto di vista [point of view], numero di telefono [telephone number]
>
> NADJ: tempo libero [free time], crisi economica [economic crisis].

As can be seen in the examples, on a scale of structural rigidity, the VN collocations represent the extreme of maximum flexibility, because they allow for the insertion of different linguistic elements between the verb and the noun (usually adjectives and adverbs, which have been underlined in the examples). They can also be considered as combinations with conventional restrictions: there is no semantic bond which makes the use of the verb *fare* [to make / do] preferable to other verbs (for example *prendere* [to take], which is used in other languages) in combination with the noun *doccia* [shower].

The other three combinations, conversely, are fixed and do not allow for the insertion of other linguistic elements within the combinations. NN and NpreN combinations, in particular, show such structural rigidity that it is possible to hypothesise their classification as compound nouns (Granger & Paquot 2008; Tanganelli 2009).

This marked difference shows how in this study different collocations have been examined, based on different POS sequences, which give rise to combinations that are, structurally, very different from one another. The data examined, therefore, constitute a very heterogeneous set in terms of their internal structure.

2.2 Raw frequency and lexical gravity

The second aspect which makes this study different from many of those which have analysed word combinations as an index for distinguishing text genres is the fact that we have used both raw frequencies and the recently proposed measure of lexical

gravity (Daudaravičius & Marcinkevičienė 2004): a statistical measure of association based on type frequency[1] as well as on token frequency. Despite this important element of innovation with respect to more classical measures, such as mutual information or log-likelihood, lexical gravity has been used relatively little to date in the study of collocations (to our knowledge only by Gries 2009, Gries, Mukherjee 2010, Ferraresi 2011; Ferraresi, Gries 2011, as well as by Daudaravičius & Marcinkevičienė 2004, who introduced it). Like other association measures, lexical gravity tries to establish the attraction strength which ties the elements of a collocation together; in order to do so, it examines the frequency of the collocation and the frequency of the individual words that compose it; lexical gravity, however, adds a new element which is also extremely important for this research: type frequency. The following is the formula for the calculation of the G-value (i.e. the value of lexical gravity):

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

In the formula, n(x) and n(y) correspond respectively to the type frequency of x and to the type frequency of y; to draw on Gries & Mukherjee (2010), there are five elements considered for the calculation of lexical gravity:

1. The frequency of the combination;
2. The frequency of word x;
3. The frequency of word y;
4. The frequency of types after x;
5. The frequency of types before y.

The value of lexical gravity grows if the frequencies of 1, 4 and 5 increase; consequently, the diversification of the words which make up the collocations plays a crucial role in the calculation of the measure of their association. Table 1 exemplifies the difference in the G-values of two VN collocations.

1 By type frequency we mean the frequency of the unique forms which constitute a text (types).

Table 1. an example of the G-values correlated to n(x) and n'(y)

	xy	n(x)	n'(y)	G-value
scendere le scale	47	1	2	-0,13
fare compagnia	47	89	2	0,51

The two collocations *scendere le scale* [go downstairs] and *fare compagnia* [keep company] have equal frequencies (47), but the elements which make up the second have a higher type frequency (the verb *fare* occurs in 89 other VN sequences with nouns other than *compagnia* (*fare parte [belong], fare la spesa [do the shopping], fare una domanda [ask a question],* etc.). The G-value of *fare compagnia,* consequently, is higher than the value of *scendere le scale.*

In section 3, we will see how the use of raw frequency and lexical gravity gives rise to different and complementary results in the analysis of the distribution of collocations across different text genres.

2.3 Data

The collocations made from selected POS sequences (VN, NN, NpreN, NADJ) have been extracted from the *Perugia Corpus*, a reference corpus of 25 million words of written and spoken Italian. The *Perugia Corpus* is made up of 10 sections which correspond to 10 different textual genres[2]; six of these have been chosen for this study; three of these are written:

- academic texts (1,100,000 tokens);
- literary prose (3,600,000 tokens);
- school essays (1,200,000 tokens);

and three are spoken:

- dialogic speech (1,000,000 tokens);
- monologic speech (1,100,000 tokens);
- film dialogues (630,000 tokens).

2 The 10 sections that compose the Perugia corpus are: academic prose, administrative texts, school essays, literary fiction, non-fiction texts, spoken texts (dialogic spontaneous speech and monologic speech), television transcriptions, web texts, press, film dialogues (http://perugiacorpus.unistrapg.it).

The collocations extracted have been filtered based on their frequency (only collocations with frequencies of >4 per million words have been selected), and then manually, in order to remove non-collocations. The final list obtained is made up of 3 762 collocations, distributed in text genres as reported in table 2:

Table 2. number of collocations for each text genre.
Highest values are in bold, lowest values are underlined

	VN	NN	NpreN	NADJ
Academic texts	303	30	**338**	**1328**
Literary prose	467	5	114	152
School essays	460	16	134	450
Dialogic speech	353	29	101	255
Monologic speech	404	**33**	184	703
Film dialogues	**489**	19	105	139

From the first data obtained, one can infer that there are evident differences in the distribution of different types of collocations in the six text genres examined; in particular, the academic prose has, by far, the highest values for NpreN and NADJ frequencies, while film dialogues have the highest values for VN frequencies. Moreover, the literary texts seem not to prefer NN collocations, which present the highest values in monologic speech.

These results, which are statistically significant (a z-test resulted in a p-value < 0.0001), thus provide an affirmative answer, although not yet in great depth, to the first research question: different collocations seem to be capable of identifying different textual genres.

3. Discussion

The frequency values alone seem to suggest that the preference for a particular POS combination could be connected to the grammatical category resulting from the combination. In other words, it seems that the text genres that have more nouns may

have a preference for noun combinations (NN, NADJ and NpreN), and that the text genres in which verbs are more frequent may, in the other direction, show a preference for the use of verb combinations (VN).

In order to verify this hypothesis we have used a correlation test, the results of which are summarised in table 3: the Pearson correlation index indicates that the correlation between the frequency of noun sequences and the frequency of nouns on one hand, and between the frequency of verb sequences and verbs on the other, is rather high, with the exception of NN sequences, which do not show any particular correlation.

Table 3. Pearson correlation index values

VN and verbs	**0.81**
VpreN and nouns	**0.89**
NADJ and nouns	**0.91**
NN and nouns	0.42

However, this correlation does not allow, in our case, for the adoption of equivalence, proposed by Halliday (1994) among others, according to whom written texts are richer in nouns since they are 'products', with a high degree of definiteness and completeness, while spoken texts are richer in verbs, because they are dynamic and more process-oriented.

The distribution by frequency of the collocations in text genres (table 2), in fact, presents data which deviate from this equivalence, highlighting in particular that:

- a written genre, literature, gives the lowest frequency in one of the three noun combinations (NN) and is among the lowest frequencies of the other two (NADJ and NpreN);
- two written genres, school essays and literary texts, have very high frequencies of verb collocations (VN);
- a spoken genre, monologic speech, gives the highest frequency of one of the noun collocations (NN) and a very high frequency of another (NADJ).

Consequently, we may assume that the preference of specific textual genres for specific POS combinations cannot be

explained exhaustively with frequency data alone and with a sharp distinction between written and spoken genres. In an attempt to answer the second research question (are there statistical measures that are particularly appropriate for this task?), we have conducted an experiment to evaluate the effectiveness of lexical gravity in the task of analysing the distribution of different types of collocations in different textual genres.

As already mentioned, higher values of lexical gravity are obtained if, within the same POS sequence, the words which compose a collocation occur in association with other, different words; in other words, if they have a high type frequency value. What we have tried to verify is whether the measure of this preference of specific words for specific POS sequences is one of the factors which determines the different distribution of collocations in different textual genres.

In order to do this, we have extracted the G-values of all 3 762 collocations included in this study, and we have subjected them to a cluster analysis. The results are illustrated in the following three graphs.

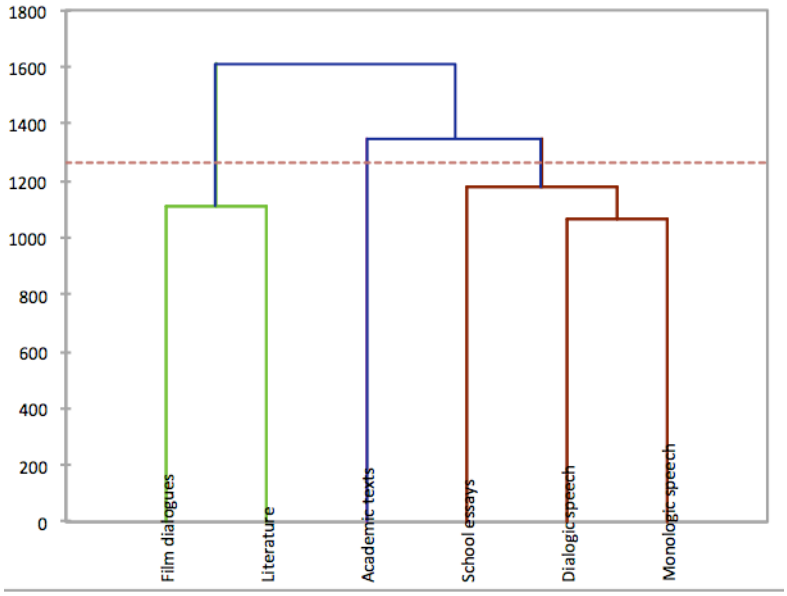

Figure 1. Cluster analysis of VN G-values

In the VN sequences (figure 1), the cluster analysis carried out on the values of lexical gravity gives rise to three classes. The first class is made up of film dialogues and literary texts, which seems reasonable since both genres have a narrative function in common. Together, speech (monologic and dialogic) and school essays form the second class; this is probably due to the fact that adolescents of a school-going age partly tend to reproduce in their writing the structures that they use in speech. Finally, academic prose makes up one single class of its own; this also seems coherent with the fact that academic texts are the only ones, among those analysed, to have a clear informative function (Biber *et al.* 2004; Spina 2010c).

Figure 2 shows the dendrogram of the cluster analysis of the lexical gravity values for the NADJ collocations; again, literary texts and film dialogues form one single class. This time, the second class includes dialogic speech and school essays, while monologic speech forms a class of its own. One explanation for this gap between dialogic and monologic speech may be the fact that monologic texts are mostly institutional speeches (lectures, court pleadings, religious sermons, parliamentary speeches), which may, therefore, differ from face-to-face speech between peers. Academic prose, again, has particular behaviour and, alone, constitutes one single class.

Finally, figure 3 shows the dendrogram of the cluster analysis of the lexical gravity values for NpreN collocations; once again, 3 of the 4 classes are formed, respectively, by narrative texts (literary and film dialogues), spoken texts (monologic and dialogic), and school essays. Again, the academic prose forms a class on its own.

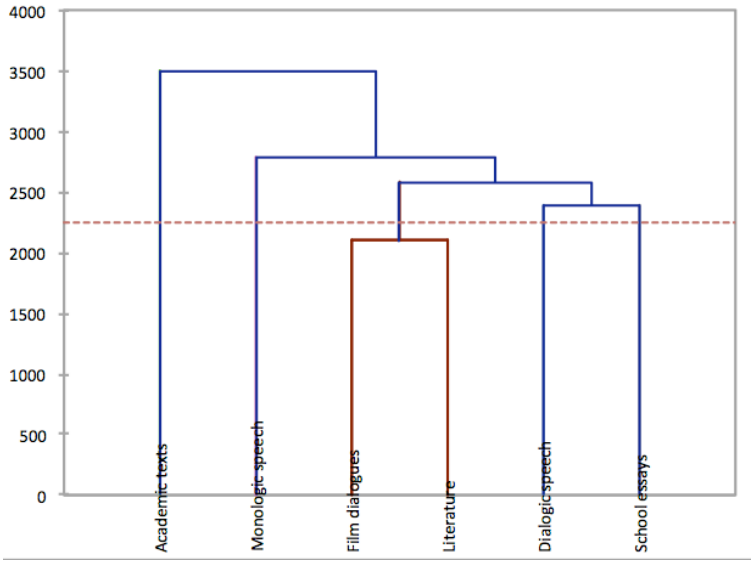

Figure 2. Cluster analysis of NADJ G-values

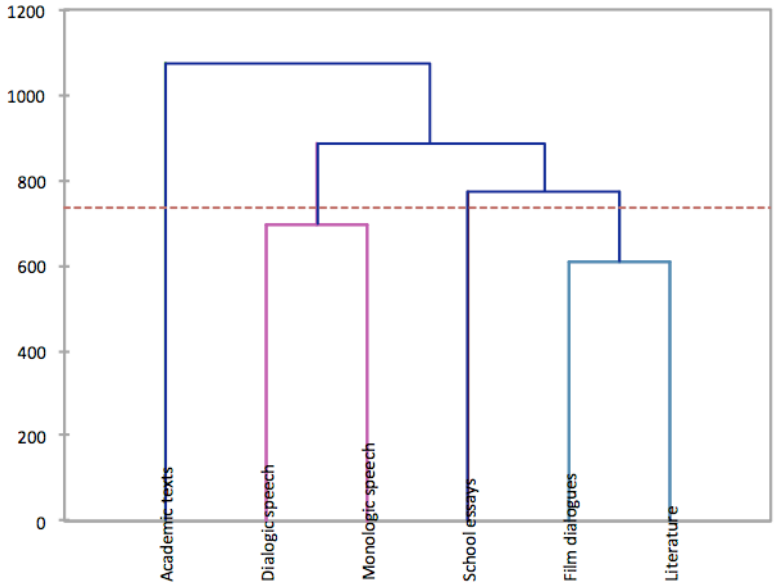

Figure 3. Cluster analysis of NpreN G-values

In light of these considerations, the data relative to lexical gravity, matched with raw frequency, seem to allow for a coherent interpretation of the variation observed in the use of different types of collocations in different text genres.

4. Conclusions

This study has sought to, firstly, verify the hypothesis according to which collocations, like other linguistic elements, can contribute to an ability to distinguish among different text genres. An analysis of the frequency of collocations formed by four different POS sequences in six genres has permitted a confirmation of this hypothesis, and an observation of some patterns: for example, that academic prose is characterised by a high number of noun collocations and by the lowest number of verb collocations observed in the study.

This information conforms with what has been demonstrated by several other studies on academic prose (Biber *et al.* 2004, for example), according to which it has a precise preference for a nominal syntax due to its predominantly informative function.

Another element which is highlighted by the frequency data of the four types of collocations is the high number of VN sequences in film dialogues. Other studies (Quaglio 2009, for example) have already demonstrated that this genre, made up of ad hoc dialogues created by screenwriters, tends to emphasise some aspects of spontaneous dialogic speech and to present it very frequently; in this case, the number of verb combinations is higher in films than in spontaneous dialogic speech.

However, frequency alone is not capable of fully explaining such a marked difference in the distribution of different POS combinations, which does not even seem coherent with respect to traditional models for the interpretation of written texts in relation to spoken ones, like the already cited Halliday (1994) model. In order to provide a coherent explanation for such variability, a lexical gravity experiment has been carried out; lexical gravity is the only measure, amongst those considered classical, which takes into account the type frequency of

the constituents in the calculation of the degree of association of a collocation.

The G-values of the 3 762 collocations that were examined and subjected to a cluster analysis allowed for the regrouping into a single class (for all four types of combinations examined) of the two genres which share the same narrative function: literary texts and film dialogues. Similarly, lexical gravity allowed us to recognize for academic prose a particular status with respect to the other five varieties, by virtue of its informative function, as previously mentioned.

In both cases, the assignment to a class through cluster analysis is not done on the basis of observing how many times collocations occur (raw frequency), but rather on the basis of observing that some words, constituents of collocations, tend to occur many more times in collocations of the same kind but in association with different words (type frequency).

As another example, the VN collocation *sentire una voce [hear a voice]* has a raw frequency in literary texts that is almost double its frequency in film dialogues (17.4 compared to 9.5 per million words); their G-values, however, are much closer than their respective raw frequencies, by virtue of the type frequency of their two components (the verb and the noun); more precisely, with regard to the verb *sentire* [hear], the decisive value is the number of other VN slots in which it occurs with other nouns.

In the case of literary texts, this number is equal to 6: the verb recurs in 6 VN sequences, together with 6 different nouns (*rumore [noise]*, *voce [voice]*, *bisogno [need]*, *odore [smell]*, *suono [sound]*, *passo [step]*). In the case of *sentire* in film dialogues, the number of other VN slots in which it occurs with different nouns is also 6 (it recurs with *voce [voice]*, *bisogno [need]*, *profumo [perfume]*, *odore [smell]*, *rumore [noise]*, *dolore [pain*]). This similarity in the syntagmatic preferences of the components within VN sequences determine, for *sentire una voce,* two lexical gravity values that are not very dissimilar (1.2 for literary texts and 0.8 for films).

The particularity of the syntactic context (the association with a noun to form a verb-noun collocation) in this example is coherent with the relative semantic and functional

contexts (the association of the verb *sentire* with nouns which denote facts that are perceptible through the senses is frequently used in the course of narrations, and, thus, it is linked to the narrative function common to both genres).

In confirmation of this, not only is the collocation *sentire una voce* absent, for example, in academic prose, but the verb *sentire* occurs in academic texts only in the VN collocation *sentire il bisogno,* which consequently has a rather low G-value (-2.18).

In conclusion, this study confirms the hypothesis that collocations can be used for the identification of text genres; together with frequency, which, like all statistical aspects linked to vocabulary (Gries & Mukherjee 2010), is strongly influenced by topic and is therefore at least partly suitable for carrying out this task, the use of lexical gravity allows the identification of regularities, depending on the text genre, in the way in which the words which make up collocations tend to recur in similar syntactic contexts, to fill the same "collocational slots", revealing a preference for particular syntagmatic contexts.

References

Biber D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber D., Conrad S. & Cortes V. (2004). « If you look at …: Lexical bundles in university teaching and textbooks », *Applied Linguistics* 25 (3): 371-405.

Biber D. (2009). « A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing », *International Journal of Corpus Linguistics* 14 (3): 275-311.

Conrad S. & Biber D. (2001). *Variation in English: Multi-Dimensional Studies*. London: Longman.

Crossley S.-A. & Louwerse M. (2007). « Multi-dimensional register classification using bigrams », *International Journal of Corpus Linguistics* 12 (4): 453-478.

Daudaravičius V. & Marcinkevičienė R. (2004). « Gravity Counts for the boundaries of collocations », *International Journal of Corpus Linguistics* 9 (2): 321-348.

Dempsey K.-B., McCarthy P.-M. & McNamara D.-S. (2007). « Using phrasal verbs as an index to distinguish text genres », *in* D. Wilson & G. Sutcliffe (eds), *Proceedings of the twentieth International Florida Artificial Intelligence Research Society Conference*. Menlo Park, California: The AAAI Press, 217-222.

Evert S. (2008). « A lexicographic evaluation of German adjective-noun collocations », *Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions*, Marrakech, Morocco.

Ferraresi A. (2011). *Exploring lexical gravity within a multi-faceted approach to the study of collocation: preliminary proposals*. ICAME 32 Conference (Oslo).

Gledhill C. (1995). « Collocation and genre analysis: the phraseology of grammatical items in cancer research abstracts and articles », *Zeitschrift für Anglistik und Amerikanistik* 43: 11-29.

Gledhill C. (2000). « The discourse function of collocation in research article introductions », *English for Special Purposes* 19 (2): 115-135.

Granger S. & Paquot M. (2008). « Disentangling the phraseological web », *in* S. Granger & F. Meunier (eds), *Phraseology. An interdisciplinary perspective*. Amsterdam: John Benjamins, 27-49.

Goldberg A.-E. (1995). *Constructions: a construction grammar approach to argument structure*. Chicago, IL: The University of Chicago Press.

Gries S.-Th. (2008). « Phraseology and linguistic theory: a brief survey », *in* S. Granger & F. Meunier (eds), *Phraseology: an interdisciplinary perspective*. Amsterdam & Philadelphia, John Benjamins, 3-25.

Gries S.-Th. (2009). « Bigrams in registers, domains, and varieties: A bigram gravity approach to the homogeneity of corpora », *Corpus Linguistics 2009 Conference*. Liverpool.

Gries S.-Th. & Mukherjee J. (2010). « Lexical gravity across varieties of English », *International Journal of Corpus Linguistics* 15 (4): 520-548.

Halliday M.-A.-K. & Hasan R. (1976). *Cohesion in English*. London: Longman.

Halliday M.-A.-K. (1994). « Spoken and written modes of meaning », *in* D. Graddol & O. Boyd-Barrett (eds), *Media Texts, Authors and Readers: a reader*, Multilingual Matters, Clevedon/Philadelphia, 51-73.

Karlgren J. & Cutting D. (1994). « Recognizing text genres with simple metrics using discriminant analysis », *Proceedings of COLING*.

Kessler B., Nunberg G. & Schütze H. (1997). « Automatic detection of text genre », *Proceedings of the 35th ACL/8th EACL*: 32-38.

Langacker R.-W. (1991). *Foundations of cognitive grammar: descriptive application.* Stanford, CA: Stanford University Press.

Laybutt B.-E. (2009). *Collocation and textual cohesion: A comparative corpus study between a genre of Written Sports Reports and a large reference corpus*. (Dissertation: http://www.asian-efl-journal.com/Thesis/Thesis-Laybutt.pdf)

Lee D.-Y.-W. (2004). *Modelling variation in spoken and written English*. London/New York: Routledge.

Louwerse M., McCarthy P.-M., McNamara D.-S. & Graesser A.-C. (2004). « Variation in language and cohesion across written and spoken registers », *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*.

Pazos-Bretaña M. & Pamies Bertrán A. (2008). « Combined statistical and grammatical criteria », *in* S. Granger & F. Meunier (eds), *Phraseology. An interdisciplinary perspective.* Amsterdam: John Benjamins, 391-406.

Peng F., Schuurmans D. & Wang S. (2003). « Language and task independent text categorization with simple language models », *in* M. Hearst & M. Ostendorf (eds), *HLT-*

NAACL 2003: Main Proceedings. Edmonton, Canada: Association for Computational Linguistics, 189-196.

Quaglio P. (2009). *Television Dialogue: The Sitcom Friends Vs. Natural Conversation*. Amsterdam & Philadelphia: John Benjamins.

Santini M. (2004). « State-of-the-art on Automatic Genre Identification », *Technical Report ITRI–04–03, 2004, ITRI*, University of Brighton (UK).

Simpson-Vlach R. & Ellis N.-C. (2010). « An Academic Formulas List: New methods in phraseology research », *Applied Linguistics* 31 (4): 487-512.

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

Spina S. (2010a). « The Dici Project: towards a Dictionary of Italian Collocations integrated with an online language learning platform », *in* S. Granger & M. Paquot (eds), *e-Lexicography in the 21st century: New Challenges, New Applications*. Presses Universitaires de Louvain, 273-282.

Spina S. (2010b). « The Dictionary of Italian Collocations: Design and Integration in an Online Learning Environment », *in* N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner & D. Tapias, *Proceedings of LREC 2010*, European Language Resources Association: 3202-3208.

Spina S. (2010c). « AIWL: una lista di frequenza dell'italiano accademico », *in* S. Bolasco, I. Chiari & L. Giuliano (eds), *Statistical Analysis of Textual Data, Proceedings of the 10th Conference JADT* (Rome, 9-11 june 2010), Editrice universitaria LED, 1317-1325.

Tanganelli E. (2009). « Lingua parlata e restrizioni lessicali. Analisi di una tipologia di poliremathiche del Lessico di frequenza dell'italiano parlato », Università per Stranieri di Perugia, http://hdl.handle.net/2447/98.