Corpus-based lexicography: an initial step for designing a bilingual glossary of lexical units in English and in Spanish

Juan-Pedro Rica-Peromingo

Keywords: *lexicography, corpus-based, phraseology, bilingual glossary, lexical units.*

Abstract

Lexicography is basically concerned with the meaning and use of words. In previous decades, lexicographers have investigated the meanings of words and synonyms, but recent lexicographic research has been extended using corpus-based techniques to study the way that words are used and, in particular, how lexical associations are used. Lexicography is, therefore, directly connected to phraseology because both disciplines study sets of fixed expressions (idioms, phrasal verbs, etc.) and other types of multi-word lexical units. This paper presents an overview of two major corpora (CEUNF and COEPROF) compiled for phraseological and lexicographical purposes: the use of lexical bundles in the writing of Spanish university students. Both the CEUNF and the COEPROF have been used to analyze the production of phraseological units (lexical bundles and grammatical collocations) present in argumentative texts written in English by Spanish EFL university students. This study, based on corpus linguistics (McEnery, Xiao & Tono, 2006), phraseology (Cowie, 1998; Howarth, 1996, 1998; McCarthy & O'Dell, 2005, Nesselhauf, 2003, 2005; Granger & Meunier, 2008) and lexicography (Atkins & Rundell, 2008; Bergenholtz et al., 2009; Hartmann, 2001, 2003; Nielsen, 2009: Ooi, 1998), uses two taxonomies taken from Biber et al. (1999) for the lexical bundles (linking and stance lexical bundles) and Benson et al. (1986, 1993) for the grammatical collocations (verbs of communication and mental processes). With these two taxonomies a bilingual list of phraseological units in Spanish and English will be devised in order to contrastively analyze the production of such units by both non-native students and professionals writing in English and with the ultimate goal of designing a lexicographical glossary of bilingual lexical units used in argumentative English writing. For the preliminary quantitative analysis of the data and word searching Wordsmith Tools (Wordlist and Collocates tools) has been used. The analysis of these initial data and the use of the appropriate statistical tools (norming of words, T-test for the statistical significance, etc.) may be seen as a starting point for producing a glossary of lexical items in argumentative writing and improved teaching material for Spanish university learners of English.

1. Introduction

Lexicography (Atkins & Rundell, 2008; Bergenholtz et al., 2009) is concerned with the meaning and use of words. Recent lexicographic research has been extended using corpus-based techniques to study the way that words are used and, in particular, how lexical associations are used. Lexicography is, therefore, directly connected to phraseology because both disciplines study sets of fixed expressions (idioms, phrasal verbs, etc.) and other types of multi-word lexical units. On the other hand, corpus linguistics has been considered by many authors as the most important methodological trend since Chomsky's revolution around 1950s. It has been extensively proved in later years that for second language acquisition (SLA) or foreign language acquisition (FLA) that the use of linguistic corpora constitutes a very useful tool for teaching languages, in order to provide our students with more practical teaching and first-hand experiences in a natural context.

But it has specially been in the field of lexicology and elaboration of dictionaries where the use of linguistic corpora has been spectacular (Ooi, 1998), in particular those dictionaries, grammar books, glossaries and reference books that nowadays take into account word frequencies, collocations and phraseology, together with variation, lexis and grammar. What has been noticeable is the fact that these new dictionaries, grammar textbooks and reference books do not only take into account native corpora, but also student corpora, which allows us to take into consideration the students' L1, and makes dictionary making not being

"market-driven", but student-driven (Hanks, 2008: 221). Therefore, we are able to provide them with more specific and appropriate teaching, since working with what the students have previously done and not with what they should, shouldn't or might do allows us to take into account "where they are, i.e. situated in their L2 learning contexts, and where they eventually (may) want to get to, i.e. close to the native-speaker language using capacity captured by L1 corpora" (Seidlhofer 2002: 215). In short, the use of student corpora allows us to analyze and compare the native and non-native students' written production.

Apart from corpus linguistics, this study has taken phraseology as its theoretical background. Phraseology (Howarth 1998; Cowie 1998; Granger & Meunier 2008) is the linguistic trend that studies lexical units, that is, the more or less free combination of terms in order to constitute units with meaning.

One of the reasons why our study is focused on phraseological units is for their wide use in most languages as a strategy for writing argumentative texts, both as adverbs (in the case of lexical bundles) or lexical phrases, to connect ideas in a logical order (i.e., *First of all*, English, *En primer lugar*, Spanish; *Zunächst*, German; *Tout d'abord*, French). A second reason for analyzing multi-word units has been our interest as university English teachers for our students' accuracy and improvement on their writing skills, since it is a strategy widely used in argumentative writing by Spanish students. And, finally, our ultimate goal is the design of a bilingual glossary of lexical units used in argumentative writing in English in order to help EFL Spanish students to improve their writing skills in English as a foreign language.

2. The study: data and methodology

This study consists of an initial analysis of the production of multi-word units which are present in English argumentative texts written by native and non-native speakers of the language. The use of multi-word units in argumentative writing has been long proved to be a basic strategy that both non-native and professional writers take hold of for their writings, specifically what Biber et al. (1999: 87) call *linking adverbials* which primarily function "to state the speaker/writer's perception on the relationship between two units of discourse" and *stance adverbials* which state the writer's position or attitude towards a unit of discourse. Some previous studies on lexical bundles (Conrad 1999; Durrant 2009; Durrant & Schmitt 2009) have found two results: first of all, most of the linking adverbials used in both conversation and academic writing are realized by single adverbs and not so much by multi-word units. Secondly, they found that phraseological units are items which are probably highly salient for native speakers. These studies will be taken into account to contrast their conclusions with the results which have been found for the study presented here.

The aims of this particular study include: firstly, to analyze the use of enumeration and addition multi-word units by non-native writers (CEUNF) with B1 and B2 levels in the university context (as stated in the *Common European Framework of Reference for Languages: Learning, teaching, assessment* 2001); secondly, to contrast their production with that done by native professional writers in English and in Spanish; thirdly, to devise a general bilingual glossary of phraseological units (lexical bundles) used in argumentative writing; and, finally, to elaborate teaching materials which enable us to find out about the use of phraseological units by non-native university writers, which will be dealt with in future papers.

Table 1. Corpora used for this study.

Corpus		Nº of words
CEUNF (*Corpus de Estudiantes Universitarios No Filólogos)*		**153.721**
COEPROF (*Corpus de Escritores Profesionales)*	**COEPROES** (*Corpus de Escritores Profesionales en Español)*	**32.584.855**
	COEPROIN (*Corpus de Escritores Profesionales en Inglés*)	**32.722.745**
TOTAL Nº OF WORDS IN ALL CORPORA		**65.461.321**

Several corpora have been used for such purposes (see Table 1 above): the CEUNF (*Corpus de Estudiantes Universitarios No Filólogos*), which is a corpus of non-native students of English from different university fields (Audiovisual Communication, Fine Arts, Computer Science, etc.) who study English as a subject outside their curriculum (Rica 2007, 2010, 2012 *in press*) and the COEPROF (*Corpus de Escritores Profesionales*, and its subcorpora, COEPROES: *Corpus de Escritores Profesionales en Español*, and COEPROIN: *Corpus de Escritores Profesionales en Inglés*), which is an original corpus of argumentative texts written by professional authors in both English and Spanish, also in different disciplines (Rica 2009).

The CEUNF (*Corpus de Escritores Universitarios No Filólogos*) corpus contains more than 150,000 words from argumentative texts written by university students at the Complutense University on current issues covered widely in the media at the time of compiling the writings, like for example: *Smoking should be banned in public places*; *Exams are not useful*; *Wars are always wrong*; *In order to study English, people need to travel, live and work in an English-speaking country*; *Mass media is of great importance in the Spanish society*, etc. These students were enrolled in English B1 and B2 level classes at the CSIM (*Centro Superior de Idiomas Modernos*) also at the Complutense University. The reason why only B1 or B2 students were selected for the corpus was mainly in order to contrast similar students' level with respect to the ICLE corpus and its Spanish subcorpus (SPICLE) in the original study previous to the one presented here. In all cases, the writers were undergraduate students majoring in fields such as Audiovisual Communication, Fine Arts, Computer Science, Biology, History, Geography, Journalism, Business studies, etc., with English being some additional training in their general studies.

The COEPROF (*Corpus de Escritores Profesionales*) contains around 65 million words from professional argumentative writings in English (COEPROIN, *Corpus de Escritores Profesionales en Inglés*) and in Spanish (COEPROES, *Corpus de Escritores Profesionales en Español*). The texts compiled for this corpus have been obtained in Acrobat format and then converted into text format. All of them are original texts published in English and in Spanish by experts in different fields (see Figure 2 below): Fine Arts, Audiovisual Communication, History, Philosophy, Psychology, etc. up to 31 disciplines in each of the subcorpora.

Since the number of words in both corpora are quite different, norming all figures by 10,000 words was applied for this study, following Biber, Conrad & Reppen (1998).

The taxonomy used has been taken from Biber (1993, 2004), Biber, Conrad & Cortes (2004) and Biber et al. (1999) for the linking lexical bundles. The reason for choosing such particular lexical patterns is because linking lexical bundles are structures commonly used in argumentative writing. Lexical bundles (specially linking adverbials) fulfil organizational and rhetorical functions which are basic in academic writing: they introduce a topic, summarize,

add information, contrast, exemplify, explain, conclude, etc. For this particular study, only enumeration and addition linking adverbials have been analyzed (see Table 2 below). Both adverbials and lexical units have been searched for.

Table 2. Taxonomy of this study.

Lexical bundles: enumeration and addition linking adverbials
● **Enumeration:** *First, Second, Firstly, Secondly, Thirdly, In the first place, In the second place, First of all, For one thing, For another thing, To begin with, Next, Finally, Lastly.* ● **Addition:** *In addition, Further, Furthermore, Similarly, Also, By the same token, Likewise, Moreover.*
Grupos léxicos: adverbiales de enlace de enumeración y adición
● **Enumeración:** *Primero, Segundo, Primeramente, En primer lugar, En segundo lugar, En tercer lugar, Antes que nada, Al principio, Desde el principio, Para empezar, Para comenzar, Como comienzo, Luego, Entonces, Finalmente, Por último, Ante todo.* ● **Adición:** *Además (de) [por otra parte], Igualmente, De la misma manera, También, De igual modo, Del mismo modo, De modo parecido, De modo similar, De esa/esta manera, De ese/este modo, Asimismo, En la misma línea.*

3. Hypotheses

For the purpose of this paper, two main hypotheses were stated: firstly, Spanish non-native students (CEUNF) were expected to use a number of phraseological units quantitatively and qualitatively different from those from COEPROES; and, secondly, the number of phraseological units used by the Spanish students writing in English was expected to be fewer than in the case of native professional writers (COEPROIN).

In order to test these hypotheses, Wordsmith Tools 3.0 (Scott, 2008) was used for the quantitative analysis of the results, while T-test has been employed for statistical significance (Moore 2007: 435), and the norming of all corpora word numbers by 10,000 in order to eliminate differences in number of words between corpora (Biber, Conrad & Reppen 1998: 263). The analysis of the initial data by using these appropriate statistical tools may allow us to, firstly, emphasize the importance of implementing multi-word units in the students' production of written texts in the university context, since it is a strategy widely used by Spanish writers in their EFL production, and, secondly, to identify transfer factors in the long term (Gass & Selinker 1983). Results will also allow us to devise an appropriate bilingual glossary with lexical units for teaching argumentation in English to Spanish university students.

4. Preliminary results and conclusions

With respect to the two Spanish corpora, preliminary results on enumeration and addition lexical bundles reveal significant differences in the use of phraseological units: the Spanish students from the CEUNF are the ones who more frequently make use of these units (see Figure 1 below).

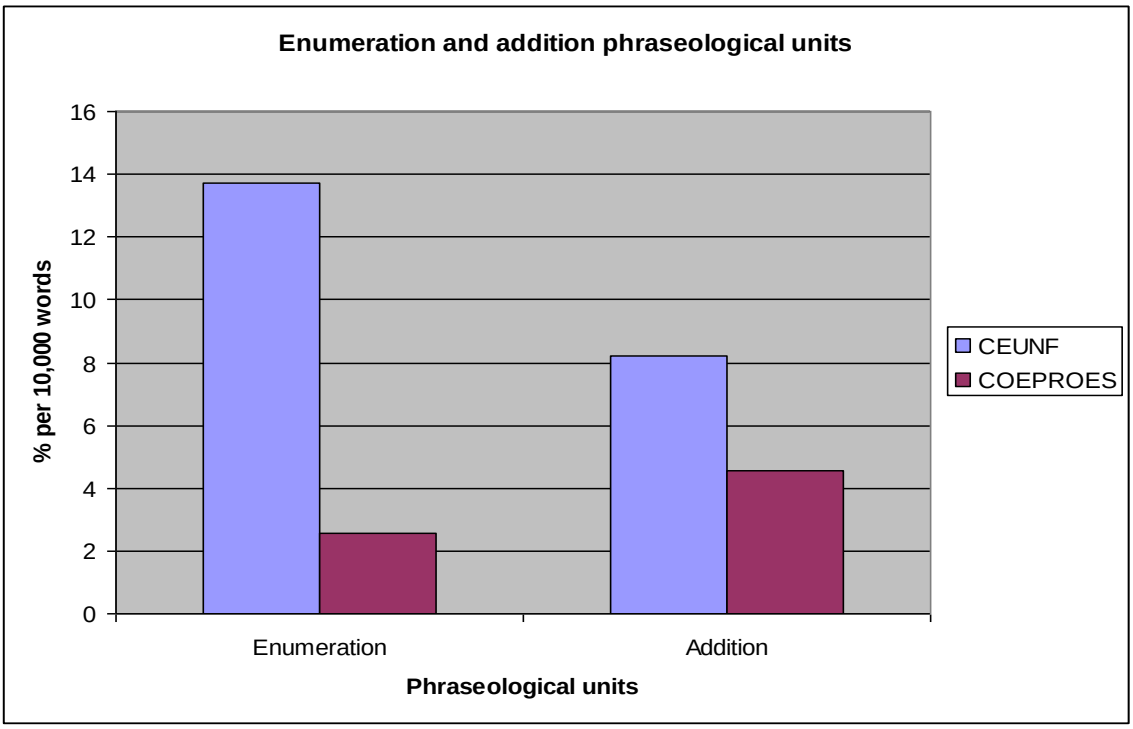

Figure 1. The use of phraseological units in CEUNF and COEPROES.

Table 3 below includes some concordance lines extracted from the two corpora which exemplify some of the uses of enumeration and addition multi-word units.

Table 3. Concordance lines: CEUNF and COEPROES.

Enumeration	**CEUNF**: 15 alike condemn all forms of self-destruction. Now let's move on to the next part which is the arguments against euthanasia. **First of all**, some people say that euthanasia is the same as assisted suicide, and that this is not the role of
	COEPROES: 7 las connotaciones peyorativas que el adusto militantismo atribuye al «esteticismo». Para precisarla, es necesario, **en primer lugar**, ir a buscar las categorías estéticas allí donde existen, allí donde viven y se inventan; es
	CEUNF: 10 We will not be able to feel wonderful and special atmosphere of live concerts. **To begin with**, I, as an intending teacher and experienced mother, am really indignant at rising violence on TV.
	COEPROES: 19 **Para empezar**, asombrará al lector enterarse de que en cierta época existían máquinas automáticas de ajedrez. En efecto, ¿cómo concebir semejantes aparatos si el número de combinaciones de las piezas en el tablero de
Addition	**CEUNF**: 34 function to musical and amusement contents, because of the likes of young people over all. **In addition**, the appearing of digital radio is an important advance because it has quite possibilities.
	COEPROES: 17 La caracteriza como la "paleotelevisión". Según él, ésta dejó su lugar a la "neotelevisión". **Además de eso**, en el artículo "TV: la transparencia perdida", analiza con detenimiento el cambio que se produce en el
	CEUNF: 11 ceremony or international summit conference. That's what many countries do. **In addition to this**, the monarchy is very expensive. We pay their cars, houses, clothes, of course, their weddings and everything else. Isn't this
	COEPROES: 151 escrita, finalmente limitados por ella, no siempre podemos comprender como tales" (Goldin, 1998:12). **En la misma línea**, Bernardo Restrepo (2002) reclamaba para la escuela una mayor dedicación al ejercicio de la palabra hablada.

Regarding the second hypothesis, and contrary to common belief, the non-native students –and not the native writers– seem to resort to enumeration and addition lexical bundles more often than native speakers of English do (see Figure 2 below), although their production is marked by an over-and underuse of certain lexical units in their argumentative writing: CEUNF students seem to overuse the multi-word unit *First of all*, whereas native professional writers tend to use a wider variety of structures: *First of all*, *To begin with* but also adverbs of the type *Lastly*, *Also* or *Likewise*, structures which are not very widely used by the non-native university writers.

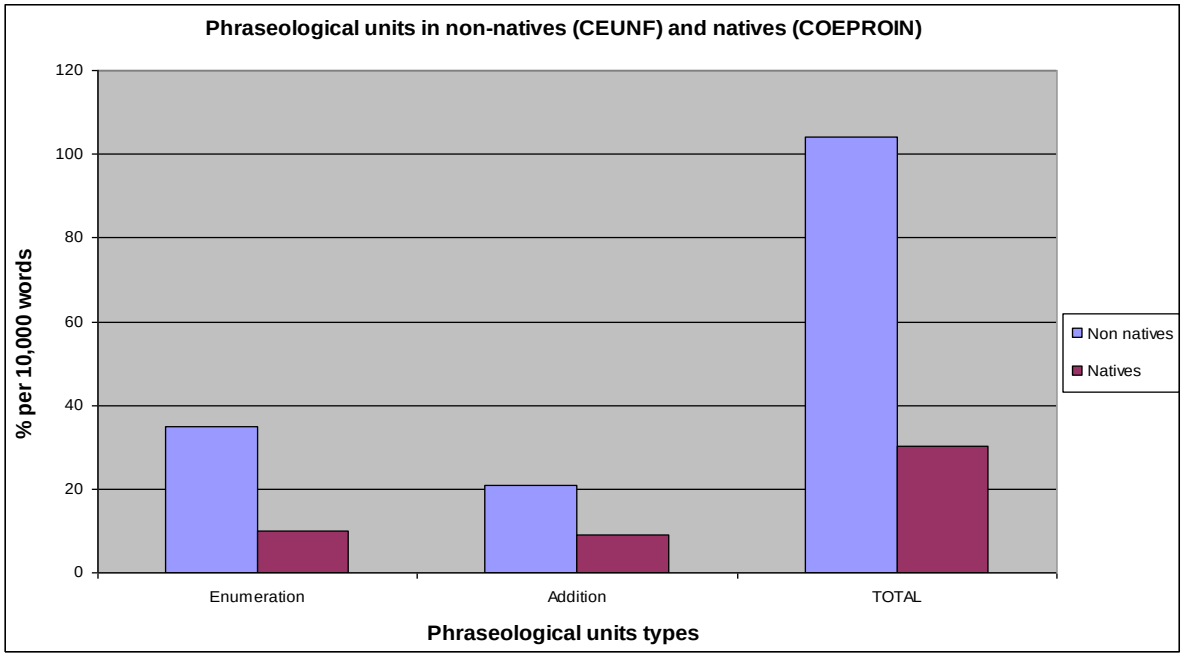

Figure 2. The use of phraseological units in CEUNF and COEPROIN.

Table 4 below includes some concordance lines extracted from the two corpora which exemplify some of the uses of enumeration and addition multi-word units and adverbs.

Table 4. Concordance lines: CEUNF and COEPROIN.

Enumeration	**CEUNF**: 61 d by *deranged* and unskilled people. Such an approach overlooks several important aspects of the gun problem in our society. **First of all**, it is important to analyze the psychological aspect of the matter,
	COEPROIN: 2 dangerous, and I am worried about what will happen." If he is worried, we all need to be worried as well. What is at stake? **First of all**, the survival of Israel . I find that my older Jewish and Israeli friends have
	COEPROIN: 12 **In the first place**, the divisions would lead one to think that it is possible to draw a sharp line, for example, between Pueblo I period and Pueblo II period. As a matter of fact, it is very difficult to make such a distinction because some elements of Pueblo I culture persisted into
	CEUNF: 33 I am going to center this composition in if smoking should be allowed or not in public places. For example, **in the first place**, in a restaurant, it is not pleasant to be eating and having to put up with the smoke
	CEUNF: 172 Sometimes children say that they are going to the room to study and they go to play computer games instead. **First**, they start with computer games and then continue with the slot machine and it is a serious
	COEPROIN: 56 Since age is positively correlated with the incidence of heart disease, we statistically adjusted the death rates for age—by taking into

	account, **first**, the median age and, second, the percentage of residents aged
	CEUNF: 2 I try to discuss the differences between these classes of school, the ways of learning and their influences on students. **Firstly**, I'll intent to comment something about the innovative or modernist technics
	COEPROIN: 45 This posture, Photo No. 22, has three main uses. **Firstly**, upon receiving a left punch to the right side of your face, you should sit backwards bringing the right foot back and swivel to your right as
	CEUNF: 3 you have to built a new country, a good opportunity for business and finances. **Lastly**, we can find two new places to do good monetary operations: El Salvador and India. The Spanish Institute for International Business (ICEX),
	COEPROIN: 27 For three hundred years the trees have been cut down faster than they could grow, first to clear the land, next for fuel, then for lumber and **lastly** for paper. Consequently we are within sight of a shortage
Addition	**CEUNF**: 31 ver our world and the decline of arts. There may be some other reasons for it, but niether science nor technology themselves. **In addition**, scientific and technological progress may very often create new
	CEUNF: 20 ntil they are older and more experienced. In other words, they cannot stick up for themselves because they are inexperienced. **Furthermore**, children cannot organize themselves in the way adults can.
	COEPROIN: 3 but it does not address the question as to exactly how these perturbations arise. **Furthermore**, a typical model of the very early universe might possess both inflationary and non-inflationary solutions,
	CEUNF: 1 because it was synonim of power, wealth and well-being. **Likewise** with smoking, it was a fashion overcoat for young people, and it was too chic smoking. The issue is that fashions comes and go away
	COEPROIN: 34 **Likewise**, the connection between the pace of life and coronary heart disease can likewise be understood as a consequence of the P-E fit. Because fast-paced places and fast-paced people both have higher

5. Conclusions

Comparing these preliminary results on enumeration and addition linking adverbials with the ones obtained by Biber et al. (1999) and Conrad (1999), important differences are found in the use of linking adverbials by non-native writers: if 80% of all the enumeration and addition linking adverbials used in Biber et al.'s study corresponded to single adverbs, in the case of the non-native writers the use of single adverbs (35,03%) and of multi-word units (35,06%) is almost the same, although there is a slightly higher number of examples of lexical units than of single adverbs. It can be affirmed, then, that the non-native students in this study rely more on multi-word units –at least with enumeration and addition lexical bundles– than those who constitute Biber et al.'s corpus. In the other very relevant study on the same issue (Conrad 1999: 9), it was found that in both conversation and academic writing, linking adverbials are represented by single adverbs and not so much by phraseological units. In this respect, our analysis does coincide with Durrant's (2009) and Durrant & Schmitt's (2009) studies, which showed that non-native writers "rely heavily on high-frequency collocations, but that they underuse less frequent, strongly associated collocations (items which are probably highly salient for native speakers)" (2009: 157).

Results also show that those phraseological units mostly used by the non-natives (and, in most of the cases, overused) are structures which are similar to the ones used in the L1 (*En primer lugar, en Segundo lugar, Para comenzar,* etc.), whereas they underuse some typical native structures found in the native texts (*For one thing, For another thing*, etc.).

A search of all types of lexical bundles in both languages included in the general bilingual list and a more qualitative study will be needed in order to confirm these initial outcomes and identify other possible factors that may influence the non-native students' writing (Levy 2008, 2010). It will also be needed in order to design the bilingual glossary with lexical units in argumentative writing that comprises our ultimate goal. As Krishnamurthy (2008: 240) states: "corpus-driven lexicography does not use a corpus to find examples to fit pre-existing entries; the new entries, sense divisions, and definitions are fully consistent with, and reflect directly, the evidence of the corpus". At the same time, the need for devising such a bilingual glossary of lexical units used in argumentative academic writing is supported by the idea that "this new view of collocation considerably widens the dictionary maker's brief, since future lexicography will have to provide a full account of both structurally simple and structurally complex units, including fixed expressions of regular syntactic-semantic composition" (Siepmann, 2005: 409).

References

Atkins, B.T.S. and M. Rundell 2008. *The Oxford Guide to Practical Lexicography*. Oxford: OUP.

Bergenholtz, H., S. Nielsen and S. Tarp (eds.) 2009. *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Berlin: Peter Lang.

Biber, D. 2004. 'Lexical bundles in academic speech and writing.' In B. Lewandowska-Tomaszczyk (ed.), *Practical Applications in Language and Computers. PALC 2003.* Frankfurt am Main: Peter Lang, 165–178.

Biber, D., S. Conrad and V. Cortes 2004. 'If you look at…: Lexical bundles in university teaching and textbooks.' *Applied Linguistics* 25: 371–405.

Biber, D., S. Conrad and R. Reppen 1998. *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: CUP.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.

Conrad, S. 1999. 'The importance of corpus-based research for language teachers.' *System* 27: 1–18.

Cowie, A.P. (ed.) 1998. *Phraseology. Theory, Analysis, and Applications*. Oxford: Clevendon Press.

Durrant, P. 2009. 'Investigating the viability of a collocation list for students of English for academic purposes.' *English for Specific Purposes* 28: 157–169.

Durrant, P. and N. Schmitt 2009. 'To what extent do native and non-native writers make use of collocations?' *IRAL – International Review of Applied Linguistics in Language Teaching* 47: 157–177.

Gass, S. and L. Selinker (eds.) 1983. *Language Transfer in Language Learning*. M.A.: Newbury House.

Granger, S. and F. Meunier 2008. *Phraseology. An interdisciplinary perspective.* Amsterdam: John Benjamins.

Hank, P. 2008. 'The lexicographical legacy of John Sinclair.' *International Journal of Lexicography* 21.3: 219–229.

Howarth, P. 1998. 'Phraseology and Second Language Proficiency.' *Applied Linguistics*, 19.1: 24–44.

Krishnamurthy, R. 2008. 'Corpus-driven Lexicography.' *International Journal of Lexicography* 21.3: 231–242.

Levy, S. 2008. *Lexical bundles in professional and student writing*. Berlin: VDM Verlag.
Levy, S. 2010. *Lexical bundles: Form, use, and function*. London: LAP Lambert Academic Publishing.
Moore, D. S. 2007. *The Basic Practice of Statistics*. New York: W.H. Freeman and Company.
Ooi, V. 1998. *Computer Corpus Lexicography*. Edinburgh: Edinburgh U.P.
Rica, J. P. 2007. *Estudio fraseológico del uso de colocaciones gramaticales y grupos léxicos en textos argumentativos nativos y no nativos: análisis de corpus de estudiantes.* Unpublished PhD Thesis. English Department I, Universidad Complutense de Madrid.
Rica, J. P. 2009. *Corpus de Escritores Profesionales (COEPROF):* COEPROES (*Corpus de Escritores Profesionales en Español*) y COEPROIN (*Corpus de Escritores Profesionales en Inglés*). Unpublished manuscript. Madrid.
Rica, J. P. 2010. 'Lingüística de corpus en la enseñanza de inglés como lengua extranjera (ILE).' In *Los caminos de la lengua. Estudios en homenaje a Enrique Alcaraz Varó*. Alicante: Publicaciones de la Universidad de Alicante, 1405–1427.
Rica, J. P. 2012, in press. 'Corpus analysis and phraseology: transfer of multi-word units.' In M. Taboada, S. Doval Suárez & E. González Álvarez (eds.), *Contrastive discourse analysis: Functional and corpus perspectives*. London: Equinox Publishing.
Seidlhofer, B. 2002. 'Pedagogy and local learner corpora. Working with learning-driven data.' In S. Granger, J. Hung & S. Petch-Tyson (eds.), *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam: John Benjamins, 213–234.
Scott, M. 2008. *Wordsmith Tools. Version 3.0. Online manual*. October 2011. http://www.lexically.net/wordsmith/.
Siepmann, D. 2005. 'Collocation, colligation and encoding dictionaries. Part I: Lexicological aspects.' *International Journal of Lexicography* 18.4: 409–443.