Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 00 (2013) 000–000

Procedia
Social and Behavioral Sciences

www.elsevier.com/locate/procedia

5th International Conference on Corpus Linguistics (CILC2013)

A Comparative Study of Collocations in a Native Corpus and a Learner Corpus of Spanish

Ana Orol González*, Margarita Alonso Ramos

Universidade da Coruña, Campus da Zapateira s/n, A Coruña, 15071, Spain

Abstract

This study proposes a method for evaluating the written production of Spanish collocations. We begin by asking if the native speaker model is the appropriate one for learners. In order to answer this question we undertook the annotation of collocations in two parallel corpora, one by native speakers, and another by learners. Once both corpora were annotated, the collocational richness of learners and native speakers were compared. In order to measure collocational richness, four parameters were established (density, variety, sophistication and number of errors). Our results show that learners do, in fact, use collocations, but their choices lack the variety, sophistication and correction exhibited by native speakers.

Keywords: collocation; lexical richness; collocational richness; lexical errors; learner corpus; native corpus

1. Introduction

The aim of this study was to delve into the nature of the problems involved in learning collocations with the intention of finding a solution that will facilitate this learning, thereby enabling learners to accomplish good collocational competence. In order to achieve this objective we compare the use of collocations in a learner corpus with the use of collocations in a native corpus. We follow the Explanatory and Combinatorial Lexicography (Mel'čuk *et al.*, 1995), which defines collocation as a restricted lexical combination that is composed of two elements (base and collocate) between which there is a particular relationship: the base selects a lexical unit, the collocate, that expresses a particular meaning as a function of the base (Alonso Ramos, 2010); for example, *paseo*

* Corresponding author. Tel.: +34 981 16 70 00; fax: +34 981 16 71 51
E-mail address: ana.orol.gonzalez@udc.es

‘walk’ selects the verb *dar* ‘give’ and *viaje* ‘travel’ selects *hacer* ‘do’ to express a similar meaning (**hacer un paseo* lit. ‘do a walk’ and **dar un viaje* lit. ‘give a walk’).

In the literature, it is generally assumed that to be fluent in a language it is necessary to know the collocations of a language (Lewis, 1993; Howard, 1998; Granger, 1998; Higueras, 2006; Ferrando, 2012) and that, consequently, the native speaker’s use of language is the model which learners have to follow in order to learn good collocational knowledge. The annotation of collocations in the two corpora has given us greater insight into the use of collocations: at first, we focused on identifying correct and incorrect collocations, but the analysis showed that the hypothetical perfection of the native speaker and the imperfection of the learner not only lies in the number of errors made and the number of collocations produced, but also in other features that we group under the term *collocational richness*, an objective measure of the appropriate use of collocations.

Until now most of the work based on corpora has focused on the learners′ production (e.g. Nesselhauf 2004), while overlooking the collocational knowledge model, native speakers. However, some studies do compare native corpora with learner corpora (Laufer & Waldman, 2011; Siyanova & Schmit, 2008; Durrant & Schmitt, 2009). These studies have broadened the knowledge of how learners learn collocations. Before describing our research, in this paper, first, we will explain two key concepts: the native speaker model and collocational richness.

As for as the native speaker model is concerned, the first question is to examine the degree to which this model must be considered the goal that learners have to achieve. The literature on language models deals with two main aspects: first, the definition of the native speaker model; second, whether or not the native speaker model is an appropriate one for learning a second language (Lee, 2006). We will not re-examine the first discussion, since we take the general definition of a language model as a reference point; but, we will follow up on the second discussion: is the native speaker model the appropriate model for learning collocations? As native speakers, we are aware that a problematic aspect in our own language is the correct and appropriate use of collocations; even if a native speaker does not usually make collocational errors, he does not always use the right word (didactic activities on lexical adequateness are included in textbooks for native speakers). However, collocations are one of the aspects which the native speaker is meant to be fluent in. Therefore, can a problematic feature be the model for learners? Is it possible that collocations define native speech while, at the same time, native speakers need to improve their use of collocations?

As for collocational richness, we have adapted this concept of lexical richness, based on statistical measures. Even if lexical richness is not always understood in the same terms by different authors[1], we have chosen the interpretation by Read (2001:200-201). For this author, lexical richness is based on the calculation of four parameters: density, variation, sophistication, and number of errors. Lexical density measures the number of different lexical words in relation to the total number of words. Variation measures the number of different words produced; sophistication the number of low-frequency words that are appropriate to the topic and style of a particular text; and the numbers of errors takes into account the total number or errors made. As Šišková (2012) points out, one of the main difficulties in assessing lexical richness is that it can only measure words as groups of letters separates by spaces, but cannot account for combinarions of words such as collocations, idioms, etc. this is the reason why we propose adapting the concept of lexical richness that assesses the production of orthographic words to the description of collocation richness that assesses the production of multiword units such as collocations.

The research questions that guide this work are the following: (i) is the native speaker model the one Spanish learners must follow in learning collocations?; and (ii) what is the distance between the learner’s collocational richness and that of the native speaker? In order to answer these questions we measured the collocational richness in two corpora. For the purpose of the present study, we used the *Corpus escrito del español L2* (CEDEL2, Lozano, 2009), which is organized thematically and by learning levels. Additionally, this has a parallel Spanish native corpus. A subcorpus of 200 texts was used to annotate collocations: 100 learner texts and 100 native texts. The

[1] Authors differ regarding what is the best formula to describe lexical richness (see Vermeer 2004; Meara & Bell 2001; Jarvis 2002; Daller *et al.* 2003; Daller et al. 2007; Malvern *et al.* 2004,). Moreover, different linguistic academic communities seem to ignore each other. Guiraud (1954), who is one of the first authors to deal with the concept of lexical richness, is the only one quoted by both the Anglophone and the Hispanic academic communities. In the Hispanic community, the concept of lexical richness has proved productive especially in relation to sociolinguistic variables, initiated mainly by López Morales (1984) (see, e.g. Ávila ,1986; Haché, 1991; Reyes Díaz, 2007-2008).

annotation process was made in two stages: first, the learner corpus was annotated (Alonso Ramos *et al.* 2010a, 2010b); and later, the native corpus. We have followed the same steps in the two stages: recognition of the collocation, identification of the elements of the collocation, assignment of the semantic and syntactic pattern, and screening of correct and incorrect collocations.

This article is organized in four sections: firstly, the four parameters that measure collocational richness and how this is calculated is explained. Secondly, the results of the calculation of collocational richness of the native and the learner corpus are analyzed and compared. Thirdly, there is a brief discussion of the particularities of collocational errors made by native speakers. Finally, we present some conclusions.

2. Establishing Parameters

Four parameters are used to measure *collocational richness* (CR): *collocational density* (CD), *collocation variety* (CV), *collocation sophistication* (CS) and *number of collocational errors* (NE). As we have already pointed out, these are an adaptation of the parameters used to measure lexical richness (Read, 2001). However, these take into account the differences between the concept of richness applied to the orthographic word compared to multiword units. Each of the parameters are explained: what is measured, how this is calculated, and, the key concepts that apply in each case.

Collocational density is used to measure the number of collocations produced in relation to the total number of words or tokens of the subcorpus. (see Table 1). The data allows us to know the total number of collocations in the subcorpus. We determine the number of collocations in relation to the number of tagged collocations, and not only in relation to the number of correct collocations in order to take into account all attempts to produce a collocation, and not only correct attempts.

Collocational variation is calculated by dividing the number of *lemma collocations* by the total number of collocations (see Table 1). We adapt the type-token ratio that traditionally measures lexical variation in order to measure the collocation token ratio. This parameter measures the number of different collocations that the learners know. Following the correspondence between lemma and inflected forms for the words, we can define lemma collocation as the base form which is the representative form of all the inflected forms of one collocation; the inflected collocation is each morphological variant of one lemma collocation. For example, *llamar la atención* 'to attract attention' is the lemma collocation of *llama la atención* '[he, she, it] attracts attention' and *llamó la atención* '[he, she, it] attracted attention', two inflected collocations.

In order to measure collocational sophistication, we start with a premise: low-frequency collocation is considered the most sophisticated, following the approach adopted by Laufer and Nation (1995) with their Lexical Frequency Profile, for whom most frequent words are easier than infrequent words. There are, however, other approaches that propose measuring lexical sophistication based on teacher judgments as well as on similarity with the L1(Tidball & Treffes-Daller, 2008). Since we opt for assuming the equivalence between sophistication and low frequency, to calculate the index of this parameter we need to obtain the frequency of the collocations of the text. We consider that a collocation is sophisticated if its base is sophisticated, that is, if its index was allocated to the lower frequency band, up to three occurrences per one million words (Almela *et al.* 2005). To assign the frequency of the base we have identified the number of occurrences of each word in a reference corpus *EsTenTen* (this corpus can be found on the *Sketch Engine* interface, Kilgarriff *et al*. 2004). We used a frequency index (from 1 to 5) based on the result of dividing the number of occurrences of the one base by the number of tokens of the corpus and, then, multiplying the result by one million. The number of lemma collocations is the divided by the total number of collocations (see Table 1).

Finally, we identify the number of wrong collocations in relation to the total number of attempts to produce collocations thereby including both correct and incorrect collocations (see Table 1). Since the nature of errors is different in native speakers' texts and in learners' texts, we have used two different typologies for their classification.

Table 1. Formula for calculating the parameters of collocational richness

Parameter	Formula

Collocational density	CD=Nº of tagged collocations / total nº of collocations
Collocational variety	CV=Nº of lemma collocations / total nº of collocations
Collocational sophistication	CS=Nº of sophisticated collocations / total nº of collocations
Number of errors	NE=Nº of mistakes / total nº of collocations

3. Collocational richness: overview of results

From the annotation of the corpus we obtain the number of collocations produced, the number of correct collocations and the number of incorrect collocations for each corpus. The learner corpus contains 1.863 collocations: 457 wrong and 1.407 correct ones. The native corpus contains 1.105 collocations: 35 wrong and 1.070 correct collocations. The number of collocations in both corpora can call the native speaker model into question: if we calculate the average number of collocations per text, learners produce eighteen collocations per text and native speakers, eleven. However, we must take into account two aspects: a) learner texts are shorter (the number of tokens is greater in learner texts than in native texts: 52.171 in learners' and 32.628 in texts by native speakers) and, as it has been claimed by e.g. Malvern and Richard (1997) the text length has an impact on the measure of lexical richness; b) collocational density (the number of collocations in relation to the number of words) is one of the parameters of collocational richness, but not the only one.

3.1. Collocational density: analysis and results

The lexical density index allows us to know the number of collocations produced in relation to the number of words produced. This is the only parameter where learners perform better than natives, although only slightly. A difference of only 0.001 (see Table 2) is negligible, therefore, we can conclude that there is no real difference.

$$CD = \frac{\text{Nº of tagged collocations}}{\text{Total number of collocations}} \tag{1}$$

Table 2. Index of collocational density

	Index of collocational density
Natives	0.034
Learners	0.035

3.2. Collocational variation: analysis and results

Natives produced 926 lemma collocations of the 1.070 inflected collocations while learners produced 824 lemma collocations of 1.406 inflected collocations. In order to know the difference between variation in native texts and learner texts, we compared the values in the index and found that there is greater variation in native speakers' texts (see Table 3). In relation to this data we can also analyze the number of collocations that are not repeated, in other words, how many lemma collocations are instantiated only once: natives produced 845 cases and learners only 637.

$$CV = \frac{\text{Nº of lemma collocations}}{\text{Total number of collocations}} \tag{2}$$

Table 3. Index of collocational variation

	Index of collocational variation
Natives	0.087
Learners	0.026

3.3. Collocational sophistication: analysis and results

Collocational sophistication is one of the parameters where we notice the most differences between natives and learners, in addition to the number of errors. We identified 93 sophisticated collocations in native texts as opposed to the 37 sophisticated collocations found in learners' texts. Table 4 shows that the sophistication index of native speakers triples the learners' index. Moreover, out of the 621 collocations used by native speakers, 88 had a sophisticated base while this occurred in 78 of the 392 collocations used by learners

$$CS = \frac{\text{Nº of sophistication collocations}}{\text{Total number of collocations}} \tag{3}$$

Table 4. Index of collocational sophistication

	Index of collocational sophistication
Natives	0.087
Learners	0.026

3.4. Number of collocations errors: analysis and results

Not unexpectedly, the number of errors is the parameter where there is the greatest difference between learners and natives since it is here where learners reveal their limitations in their command of collocational knowledge. Natives only made 35 errors as opposed to the 457 that learners made. This is the reason for the marked difference between the two indices which we can see in Table 5. Even with this considerable difference, it is surprising that natives make some collocational errors. In section 4 we discuss the typology of native errors in comparison with that of learners' errors.

$$errors = \frac{\text{Nº of mistakes}}{\text{Total numbers of collocations}} \tag{4}$$

Table 5. Index of number of errors.

	Index of number of errors
Native	0.032
Learners	0.32

3.5. Calculation collocational richness

Collocational richness is obtained by a simple formula (see Figure 5). In order to determine the mathematical functions, we have to take into account the positive and negative values of the parameters. Variation, density and sophistication are positive parameters: more variation, more density and more sophistication involve greater collocational richness; however, a greater number of errors results in less collocational richness. These three positive values are added, and, finally, the number of errors is subtracted from this result (see 5).

(5)

$$\left(\mathrm{CV} + \mathrm{CD} + \mathrm{CS}\right) - \mathrm{NE} = \mathrm{CR}$$

In Table 6 we can see that the values for the native speakers' collocational index are higher than those that indicate learners' collocational richness. Therefore, we can say that the native speaker model is still a valid model.

Table 6. An example of a table.

	CD	CV	CS	NE	CR
Natives	0.034	0.086	0.087	0.032	0.949
Learners	0.035	0.58	0.026	0.032	0.321

4. Typology of native speaker errors

After the first analysis, we noticed that not only do learners have problems with collocations, but that natives do also, although they have different types of problems. Interference is the main cause of error: interlingual in learners and intralingual in natives. Learners make errors such as *gastar el tiempo lit. 'spend the time' instead of perder el tiempo lit. 'lose the time', as a literal translation of to spend time (Alonso Ramos et al. 2010a, 2010b). Native speakers' errors are produced by interferences with their own language: a speaker produces an incorrect collocation, but has a correct collocation in mind which has been mixed with another correct collocation («greffes collocationnelles» in Polguère 2007).

The number of native speakers' errors is small in relation to the number of learners' errors. However, it is possible to classify them according to a typology that can be used to explain the problems learners have when producing collocations. The next figure shows a diagram of the different types of native speakers' errors (see Figure 1).

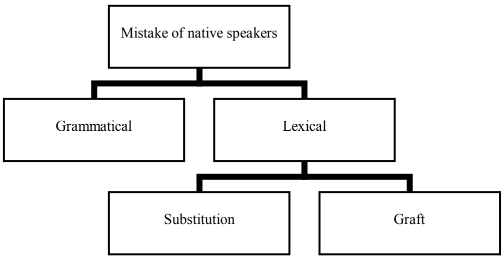

Fig. 1. Mistakes of native speakers

Grammatical errors concern number, gender, government, etc. For example, *andar de bicicleta lit. 'walk of bicycle' by andar en bicicleta lit. 'walk in bicycle', 'ride a bicycle'. Two types of errors can be distinguished: substitution errors and collocational grafts. Substitution errors can be produced for two reasons: first, the incorrect choice of one of its elements such as the case of *echar marcha atrás lit. 'pour step back' by dar marcha atrás lit. 'give step back'; second, the use of a correct collocation, but with a different sense than what the speaker wants to say, for example, salud global 'global health' for expressing buena salud 'good health'.

Collocational grafts are another kind of error which is more complex. Polguère (2007) explains grafts (greffes collocationneles), as a particular case of native collocational error. Grafts can be defined as the production of an incorrect collocation as a result of the interference of two collocations, both correct: the one that we have in mind; the other that interferes in the production. Three collocations take part in the process: the wrong collocation (graft), the collocation that the speaker wants to produce (target collocation) and the collocation where the interference is produced (source collocation). For example, the case of *llamar la intención lit. 'call the intention', incorrect production; the target collocation would be tener la intención lit. 'have the intention' and the source collocation would be llamar la atención lit. 'call the attention'.

5. Conclusion

The contrastive study between a native speaker corpus and a learner corpus that we have carried out provides data that allows us to more thoroughly understand the difficulties of learning collocations for learners. Moreover, the results point the way towards the elaboration of useful guidelines that will make learning collocations easier. On the one hand, the work with corpora has allowed us to reject some common assumptions. For instance, we have proved that learners do, indeed, use collocations, but their choices do not have the degree of variety, sophistication and correction of native speakers. Therefore, some assumptions such as the notion that learners use few collocations are shown to be false. On the other hand, the parameters that define native speakers' collocational richness indicate that error is the key factor in terms of producing correct collocations. This could certainly be taken as an incentive for placing greater emphasis on the relevance of collocations in didactic material.

Acknowledgements

This work has been supported by the Spanish Ministry of Economy and Competitiveness (MINECO) and the FEDER Funds of the European Commission under the contract number FFI2011-30219-C02-01.

References

Alonso Ramos, M. (2010). No importa si la llamas colocación o no, descríbela. In C.Mellado, C. et al. (Eds.), *La fraseografía del S. XXI: Nuevas propuestas para el español y el alemán*, Frank & Timme, Berlin, 55-80.

Alonso Ramos, M., Wanner, L., Vázquez, N., Vincze, O., Mosqueira, E., & Prieto, S. (2010a). Tagging collocations for learners. In S. Granger, M. Paquot (Eds.), *eLexicograpy in the 21st century: New Challenges, New Applications. Proceedings of eLex 2009*, Cahiers du Cental 7, Louvain-la-Neuve, Presses universitaires de Louvain, 369-374.

Alonso Ramos, M., Wanner, L., Vincze, O., Casamayor, G., Vázquez, N., Mosqueira, E., & Prieto, S. (2010b). Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora. In *7th International Conference on Language Resources and Evaluation (LREC)*, La Valetta, Malta, 3209-3214.

Almela, R., Cantos, P., Sánchez, A., Sarmiento, R., & Almela., M. (2005). *Frecuencias del español. Diccionario y estudios léxicos y morfológicos*. Madrid. Universitas, S.A.

Ávila, R. (1986). Léxico infantil de México: Palabras, tipos, vocablos. In *Actas del Congreso del II Congreso Internacional sobre el español de América*, México, D.F.: Universidad Nacional Autónoma de México, 510-517.

Daller, H., Milton, J., & Treffers-Daller, J. (Eds.) (2007). *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in spontaneous speech of bilinguals. *Applied Linguistics* 24: 197–222.

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

Ferrando Aramo, V. (2012). *Aspectos teóricos y metodológicos para la compilación de un diccionario combinatorio destinado a estudiantes de ELE*. PhD dissertation, Tarragona, Universitat Rovira i Virgili.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.) *Phraseology. Theory, Analysis, and Applications*. Oxford, Clarendon Press, 145-160.

Guiraud, P. (1954). *Les caractéristiques statistiques du vocabulaire*. Paris: Presses Universitaires de France.

Haché, A. M. (1991). Aportes de las pruebas de riqueza léxica a la enseñanza de la lengua materna. In H. López Morales (ed.), La enseñanza del español como lengua materna, Río Piedras: Universidad de Puerto Rico, 47-60

Higueras, M. (2006). Las colocaciones y su enseñanza en clase de ELE. Madrid: Arco Libros.

Howarth, P. (1998). The phraseology of learners. In A.P. Cowie (Ed.) Phraseology. Theory, Analysis and Applications. Oxford: Oxford University Press, 161-186.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004).The Sketch Engine. Proceedings of Euralex 2004 , 105–116. Lorient, France.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. Language Testing, 19: 57–84.

Laufer, B. & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. Applied Linguistics. 16 (3), 307-322.

Laufer, B., & Waldman, T. (2011). Verb-Noun collocations in second language writing: a corpus analysis of learners' English. Language Learning, 61, 647-672.

Lee, J. (2005). The native speaker: an achievable model? The Asian EFL Journal Quartely, 7.

Lewis, M. (2000). Teaching collocation. Further developments in the lexical approach. London: Language Teaching Publications.

López Morales, H. (1984). La enseñanza de la lengua materna. Lingüística para maestros de español, Madrid: Editorial Playor.

Lozano, C. (2009). CEDEL2: Corpus Escrito del Español L2. In C.M. Bretones Callejas et al. (Eds.) Applied Linguistics No: Understanding Language and Mind / Lenguaje y la Mente. Almería. Universidad de Almería, 80-93.

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). Lexical Diversity and Language Development: Quantification and Assessment. Houndmills, Basingstoke: Palgrave Macmillan.

Malvern, D.D., & Richards, B.J. (1997) A new measure of lexical diversity. In Ryan, A. and Wray, A., (Eds.), Evolving models of language. Clevedon: Multilingual Matters, 58–71.

Meara, P., & Bell, H. (2001). P_Lex: a simple and effective way of describing the lexical characteristics of short L2 texts. Prospect 16: 5–19

Mel'čuk, I., Clas, A., & Polguère, A. (1995). Introduction à la lexicologie explicative et combinatoire. Louvain-la-Neuve: Duculot.

Nesselhauf, N. (2004). Collocations in a Learner Corpus. Amsterdam: John Benjamins Publishing Company.

Polguère, A. (2007). Soleil insoutenable et chaleur de plomb : le statut linguistique des greffes collocationnelles. In M.-C. L'Homme and S. Vandaele (dir.), Lexicographie et terminologie : compatibilité des modèles et des méthodes, Les Presses de l'Université d'Ottawa, Ottawa, 247–291.

Reyes Díaz, M. J. (2007-2008). Riqueza léxica de textos redactados por alumnos de bachillerato de Las Palmas de Gran Canaria. Anuario de Lingüística hispánica 23-24: 147-163.

Tidball, F., & Treffers_Daller, J. (2008). Analysing lexical richness in French learner language: what frequency lists and teacher judgements can tell us about basic and advanced words. French Language Studies, 18, 299-313

Siyanova, A., & Schmitt, N. (2008). L2 Learner production and processing of collocation: a multi-study perspective. The Canadian Modern Language Review, 64, 3, 429-458.

Šišková, Z. (2012). Lexical Richness in EFL Students' Narratives. Language Studies Working Papers. 4, 26-36.

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards and B. Laufer (Eds.), Vocabulary in a Second Language: Selection, Acquisition and Testing, Amsterdam/Philadelphia, John Benjamins, 173,189

Tweedie, F.J., & Baayen, R.H. (1998). How variable may a constant be? Measures in lexical richness in perspective. Computers and the Humanities 32, 323-352.