Comparing learners' and native speakers' use of collocations in written Spanish[1]

Marcos García Salido
marcos.garcias@udc.gal
Marcos Garcia
marcos.garcia.gonzalez@udc.gal
Universidade da Coruña
Lengua y Sociedad de la información (LyS) research group
Departamento de Letras

Abstract: This article compares the use of collocations in texts written by native speakers and advanced learners of Spanish. The collocations studied were first identified in the sample texts on the grounds of phraseological criteria and subsequently assigned frequency information corresponding to their occurrence in a reference corpus of Spanish. The distribution of collocations with different frequency of co-occurrence and mutual information scores in the texts of the two samples was then compared, as was their proportion in the collocation repertoire of native speakers and learners. The study revealed significant differences both in terms of frequency and mutual information.

Key words: collocations, learner writing, association measures, frequency, mutual information

[1] This research was supported by two post-doctoral grants from the Xunta de Galicia (POS-A/2013/191) and the Spanish Ministry of Economy, Industry and Competitivity (Juan de la Cierva formación FJCI-2014-22853).

1 Introduction

It has been repeatedly observed that a substantial part of the linguistic production of speakers/writers consists of conventionalized combinations of lexical items (Erman and Warren, 2000; Pawley, 1985). In fact, one of the challenges non-native speakers have to face when learning a foreign language is acquiring the command of such combinations, which pose problems even for advanced learners (see, for instance, Altenberg and Granger, 2001; Gilquin, 2007; or Vincze et al., in press). Mastering this type of combinations not only gives a native-like quality to learners' production, but also seems to be related to fluency and hearer comprehension (Wray 2002: 145). Collocations are viewed as a prominent subset of this lexico-combinatorial repertoire, even though the term is far from being univocal. There seem to be two fundamental ways of understanding this concept. The first of these, the origins of which can be traced back to Firth (1957[1951]) via Halliday (1966),[2] stresses the importance of frequency of co-occurrence as a defining trait of collocations. In this vein, Sinclair defines the collocates of a given COBUILD headword as "lexical items occurring within five words either way of the headword with a greater frequency than the law of averages would lead you to expect" (Sinclair, 1987: 70).

The second approach, usually linked to the Russian lexicological tradition (see Cowie, 1981: 226-227), understands collocations in a phraseological way – frequency of co-occurrence is more or less irrelevant in defining collocations under this conception. In contrast to the previous approach, lexical semantics plays a more prominent role, to the extent that for some authors within this framework the key feature of the notion is that the meaning of one of the lexical items of the collocation is determined by the presence of a co-occurring lexical element (Cowie: 1981: 227, 1998: 215; Hausmann, 1989: 1010). From this perspective, in Spanish the verb *dar* could be considered to be a synonym for the verb *causar* 'to cause' only when it takes certain

[2] Halliday (1966) takes up Firth's idea of studying lexis at a level of analysis independent of lexical semantics ("Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words" [1957 [1951]: 195]) and, additionally, operationalizes the concept of collocation in terms of frequency in a way similar to Sinclair's: "In a lexical analysis it is the lexical restriction which is under focus: the extent to which an item is specified by its collocational environment. This therefore takes into account the frequency of the item in a stated environment relative to its total frequency of co-occurrence. [… T]here will be environments such that a string occurs with a probability greater than chance" (Halliday, 1966: 156). For a more thorough review of the history of the Firthian approach the reader is referred to Krishnamurthy (2000) or Vincze (2015: 6–11).

objects (*dar miedo=causar miedo* 'to give, cause fear; to frighten'; in contrast to *dar una patada≠??causar una patada,* 'to give a kick' or **dar un accidente ≠ causar un accidente* 'to cause an accident'). Igor Mel'čuk can be considered a representative of this tradition, but the theoretical framework he proposes, the Meaning-Text Theory (MTT), places a special emphasis on linguistic production. Thus, within this particular framework, in the previous example, rather than saying that the meaning of *dar* is contextually conditioned by *miedo*, one should state that, if the speaker wants to convey the sense 'to cause' in the context of miedo, s/he is constrained to use, among other possibilities, the lexeme <*dar,* 'to cause'>, rather than, for instance <*hacer*, 'to cause'>, a valid choice in other contexts (e.g., *hacer gracia,* 'to cause amusement'). The type of restriction at play here is, therefore, a lexical one, since it is the choice of a particular lexeme expressing the meaning 'to cause' that is not free. Moreover, within MTT, collocations are conceived of as binary combinations of elements in an asymmetric relationship, whereby one lexeme guides the choice of the other. In what follows we will adopt this conception and refer to the guiding lexeme as the *base*, and the one chosen depending on the former as the *collocate* (see Mel'čuk, 2012: 39).

This study will compare the distribution of collocations in the texts of native speakers and learners of Spanish, and to this end makes use of analytical tools taken from the two approaches discussed above. Following the second of these, the sample of collocations we examine was manually identified according to phraseological criteria and extracted from sections of the *Corpus escrito del español como L2* (henceforth CEDEL2; Lozano, 2009; Lozano and Mendikoetxea, 2013). Additionally, this set of so-called "phraseological collocations" has been assigned two association measures often used in research informed by the first approach, namely frequency and mutual information, obtained from a large corpus of general European Spanish (*esTenTen11;* Kilgarriff and Renau, 2013). Our main aim is to classify the collocations found in a sample of learners' and native speakers' essays according to these association measures and examine (i) how the groups resulting from this classifications are distributed in the texts of the two aforementioned populations and (ii) which proportions of these populations' collocation repertoire they represent. As will be seen in Section 2, similar research has been carried out for English and the results are revealing with respect to the types of collocation — in terms of their frequency and other measures derived therefrom — that are more easily processed or more conspicuously used by learners and those which are underused by them.

The article is structured as follows. The following section (Section 2) reviews some of the research studying learners' phraseology by means of association measures. Section 3 describes the methodology used in the present study. Its results are presented in Section 4 and discussed in Section 5, before moving on to the conclusions (Section 6).

2. Association measures and the use and processing of lexical combinations

In the last few years several studies have examined the use made by native and non-native speakers of combinations of lexical items in terms of measures that quantify the degree of association of such combinations (Lorenz, 1999; Bestgen and Granger, 2014; Durrant, 2008; Durrant and Schmitt, 2009; Granger and Bestgen, 2014). Other scholars (Ellis et al., 2008) have also investigated the relation between the degree of association of lexical items and the processing of the resulting combinations. This section examines the association measures most frequently used and the results of the cited studies.

2.1. Frequency, *t-score* and mutual information

In spite of the fact that the repertoire of association measures is relatively large (cf. Evert, 2004), a few of them seem to have sufficed to capture interesting differences between native and non-native speakers, viz. *frequency of co-occurrence*, *t-score* and *mutual information* (MI). Out of the three measures, *frequency of co-occurrence* is the most straightforward, being a simple tally of the number of times the constituents of a collocation occur together within a given span or a given syntactic configuration.

T-score, on the other hand, aims to determine whether the combination of two items in a particular context (usually a given span) is significantly more frequent than the overall frequency of its two members in the corpus at issue would lead us to expect. *Expected frequency* is calculated by multiplying the frequencies of the members of the combination (f_a, f_b) and dividing the product by the total size of the corpus[3] ($f_a \times f_b/n$) and represents the frequency we would expect assuming that the words of our corpus had been arranged in a random fashion (see Evert, 2008). Once we have obtained the observed (O) and expected frequencies (E) of a given combination of lexical items, we can calculate the t-score according to the following formula:

[3] This way of computing the expected frequency is the one used in the Sketch Engine's word sketches (see Lexical Computing, 2015). However, Evert (2004, 2008) argues for calculating the expected frequency differently depending on the way lexical collocation candidates have been extracted. Thus, if words related by syntactic dependencies of the form X → Y are the candidates for collocations, the denominator of E should be the total count of dependencies X → Y, rather than the whole corpus size, and for the particular collocation A → B the numerator should be the product of the frequency of A in the context A → Y by the frequency of B in the context X → B.

$$t = \frac{O - E}{\sqrt{O}}$$

Thus we obtain a standardised difference between the two frequencies, which can be negative or positive. A negative one would indicate that the two members of the combination repel each other, since their frequency is lower than expected if they co-occurred randomly. T-scores above 2 have been considered as an indication of significantly frequent co-occurrence (Durrant and Schmitt, 2009; Stubbs 1995). It has also been noted that the t-score measure highlights frequent collocations (see, for instance, Stubbs, 1995). This can also be verified in our sample (for further details, see Section 3.1 below): the following plot shows a strong positive correlation between frequency and t-score values.

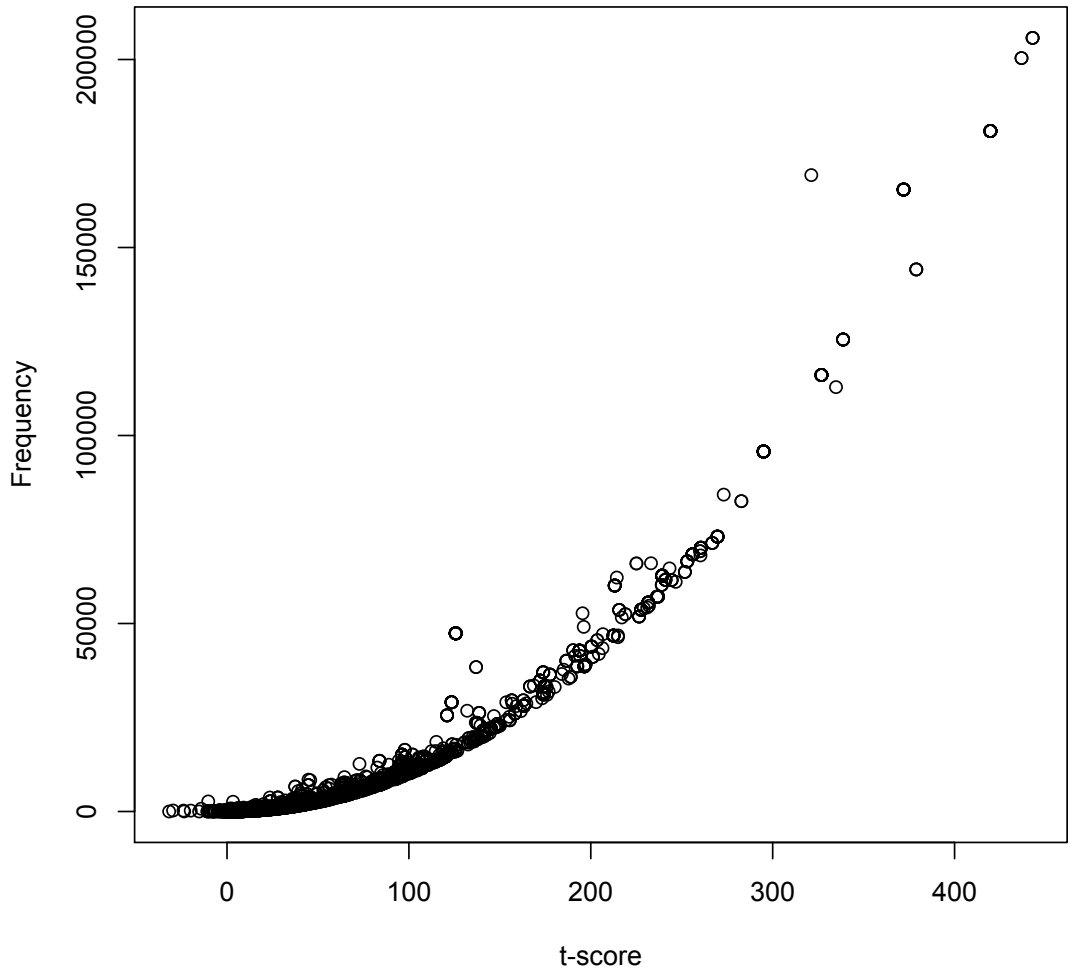


Fig. 1. Correlation between frequency of co-occurrence and t-score (see Section 3.1. for a description of the sample).

The existence of this positive correlation between t-score and frequency is also confirmed by applying Kendall's tau-b test: τ_b=0.927 ($p<0.001$).

Another widely-used association measure is *mutual information* (MI). MI is the base-2 logarithmic transformation of dividing the observed frequency of a given combination by its expected frequency, calculated in the same way as for the t-score:

$$MI = \log_2 \frac{O}{E}$$

An MI of 0 would indicate that a given combination is as frequent as randomly expected ($\log_2 1 = 0$), while an MI of 1 would suggest that the combination is twice as frequent as randomly expected, etc. As in the case of the t-score, it is possible to obtain negative values, interpretable as an indication of a certain avoidance of a given combination. One of the problems that have been noted in the case of MI is that it highlights extremely infrequent combinations (see Evert, 2008), this being the reason why it is often used in combination with a certain frequency threshold (usually ≥5). Several authors have pointed out the different rationales behind t-score and MI: the first is a test measuring the amount of evidence against the fact that a certain frequency of co-occurrence is the outcome of chance (Durrant, 2014: 455); the second (taken from Information Theory) captures the amount of information shared by the items making up a given combination (Evert, 2008). In this respect, Durrant (2008: 82) notes that "when we encounter one part of a word pair which has a high mutual information score, we can predict that the other part of the pair is likely to be nearby". MI has been praised for its intuitiveness (Evert, 2008) and high values of this measure have been regarded as an indication of coherence (Ellis et al. 2008: 380) and even idiomaticity (Lorenz, 1999: 184).

2.2. Previous studies

As stated before, some researchers have applied these association measures to the study of the use and processing of lexical combinations. One of the first to do so was Lorenz (1999), who focused on adjective intensification patterns in learners of English using only frequency data derived from the learner and native speaker corpora he analysed. According to his data, learners over-used high-frequency modifiers (adverbs and phrases; Lorenz, 1999: 185) whilst native speakers, by contrast, tended to use a larger repertoire of infrequent items. As for MI, the mean score of native speakers was higher

than that of learners, but, according to Lorenz, this measure should be interpreted in different ways depending on the group in question: in the case of native speakers a high MI score highlighted idiomatic sequences, whereas in the case of learners, it could be the result of an infelicitous combination of rare forms (Lorenz, 1999: 185).

In contrast to the method followed by Lorenz, the most usual approach for studying the distribution of combinations with different association measures in learner corpora seems to be to obtain such measures from a large reference corpus and only then apply them to the corresponding combinations from learners' and native speakers' samples (see Granger and Paquot, 2012: 135). Taking this approach, Durrant (2008: 165-182) and Durrant and Schmitt (2009) compared the distribution of premodifier+noun sequences in native and learner writing in terms of the t-score and MI score they have in the *British National Corpus*. After establishing that high t-score combinations occurred more frequently in the learner sample, whereas combinations with high MI values were used more sparsely by this group, they concluded that what distinguishes learner from native writing is not so much the absence of highly frequent phrases in the learner texts (in fact, they use them frequently), but the scarce presence of "lower-frequency but strongly associated items" (Durrant, 2008: 182; see also Durrant and Schmitt, 2009: 174).

Granger and Bestgen (2014) adopted a similar approach, although with certain modifications. First, they limited themselves to learner production. Second, they carried out a longitudinal study. Furthermore, the range of bigrams taken into account was considerably expanded by adding an adverb+adjective category to that of premodifier+noun (which they additionally split into noun+noun and adjective+noun) and by considering the whole set of bigrams in their corpus. The pattern that emerged from their study was that, as learners' proficiency level rose, their phraseological repertoire changed by increasing the share of bigrams with high MI scores while reducing the proportion of high-frequency sequences (Granger and Bestgen, 2014: 240).

The same authors (Bestgen and Granger, 2014) carried out another study that combined the longitudinal and the pseudolongitudinal approaches and paid attention, on the one hand, to the development of phraseology – operationalized again by means of t-score and MI – and, on the other, to the relation of the two above-mentioned measures with writing quality. Their results showed that t-score means decrease significantly as learners' proficiency increases. By contrast, MI turned out to be a good predictor of essay quality, as assessed by professional raters.

As for the processing of word combinations by native and non-native speakers, Ellis et al. (2008) carried out a study comprising several experiments (measuring reaction times in recognition tasks, voice onset and articulation times). Overall, MI was a significant predictor for native speakers, whereas frequency of co-occurrence was significant for learners. The authors conclude that non-native speakers "are starting to recognize and become attuned to more frequent word sequences, but they need help to recognize […] distinctive formulas [associated with high MI scores]" (Ellis et al., 2008: 391). They interpret that sequences with high MI values have clearly distinctive meanings or discourse functions and are incorporated as a whole by native speakers.

In what follows, we will examine the collocations present in essays written by learners of Spanish following a methodology somewhat similar to that employed by Durrant and Schmitt (2009), Bestgen and Granger (2014) and Granger and Bestgen (2014). As these authors did,, we will study these combinations by classifying them according to the association measures they show in a reference corpus. In contrast to the cited studies, however, our sample consists of collocations defined according to phraseological criteria, rather than of bigrams filtered by means of statistical measures. A further difference is that, as far as collocation tokens are concerned, we focus on their distribution in texts rather than on their proportion within the whole set of collocations in the sample.

3. Methodology

In this section we describe the samples of learners and native speakers we have used in the present study and the procedure followed to assign corpus frequency to the collocations of these two samples.

3.1. Samples

The collocations compared in this study come from two samples of the *Corpus escrito del español como L2* (henceforth *CEDEL2*; Lozano, 2009; Lozano and Mendikoetxea, 2013), a corpus of essays written by learners with English as L1 currently consisting of over 750,000 words. Our sample was limited to a group of texts written by learners who achieved the highest scores (ranging between 75% and 100%) in the placement test administered to them when the corpus was compiled. The study focuses, therefore, only on the writing of learners that could be considered to be of intermediate-advanced and advanced level and is of a cross-sectional (i.e., non-longitudinal) nature.

A similar number of texts was used for both native speakers and learners (ca. 100), although their mean length differed from one group to another: contrary to all expectations, it was the native speakers who produced the shortest texts. The composition of this sample can be seen in Table 1.

Table 1. Corpora composition

	N of texts	Text mean length (in words)	Standard Deviation	Size (in words)
Native Speakers	104	286.16	140.5	30,037
Learners	100	464.2	113.83	46,420

Collocations present in the two samples were manually annotated. Each text was first annotated by two native speakers of Spanish, after which the two resulting annotations were merged by a consensus annotator (for further details, see Vincze et al., 2011). In the case of the learners' sample, collocations deemed incorrect by the annotators were also annotated and classified according to an error typology proposed by Alonso Ramos et al. (2010a and b). The criteria employed by the annotators in identifying collocations were derived from the Explanatory Combinatorial Lexicology (ECL) theoretical framework (see, for instance, Mel'čuk, 2012). The collocations, identified according to these criteria, covered a range of different syntactic relations. For the purposes of this study, we have limited ourselves to the following four types:

a) verb+object: including direct and prepositional objects, e.g., *dar la bienvenida* 'to welcome', *asistir a la universidad* 'to attend university'
b) subject+verb: e.g., *sale el sol* 'the sun rises', *sube la deuda* 'debt increases'
c) noun+noun: generally structures involving a quantifier noun governing a prepositional phrase including the quantified noun, as in *paquete de tabaco* lit. 'packet of tobacco', 'cigarette packet', but also nouns modified by a classifier prepositional phrase – *película de terror* 'horror movie', or appositional structures such as *carril bici* 'bike lane'
d) noun+adjective: e.g., *día festivo* 'holiday', *larga distancia* 'long distance'

The sample included a total of 1052 collocations in the case of the native speakers' group and 1719 in the case of the learners'.

3.2. Assigning frequency and MI scores to collocations

Each collocation was assigned the frequency of co-occurrence it displayed in the *esTenTen11* corpus of European Spanish (Kilgarriff and Renau, 2013), which is lemmatised and PoS-tagged with FreeLing[4] and contains over 2 billion words. The frequency measures were obtained by means of a procedure similar to the Sketch Engine's word-sketches (Lexical Computing, 2015), namely by attending to the frequency of the two collocation members in a given grammatical pattern instead of by considering their frequency within a given span. However, we did not use the sketch-grammars available for Spanish in the Sketch Engine, but a different set of lexico-syntactic patterns developed by Vincze and Alonso Ramos (2013). The authors claim that this set of patterns improves the recall of the available sketch-grammars, as they include syntactic relations that could hardly be retrieved using such grammars, such as preverbal relativized objects (e.g., *el paseo que dio*; 'the walk s/he took'). These adapted grammars were fed with the lemmas of each constituent of the collocation.

Some of the searches were manually revised in cases that pose a particular difficulty, such as participial forms with an adjectival function (e.g. *mes pasado* 'last month') treated as adjective lemmas in our queries, but grouped under the verb lemma in the corpus (*pasado* ← *pasar*). In such cases, the lemma was replaced by a pattern-matching query that included the inflectional variants of the participial adjective in question. A similar case is that of collocations including a preposition, for which we manually checked that the query included the preposition actually employed in the source text — even if that meant, in the case of some collocations used by learners, looking for a "wrong" preposition likely to be less frequent than its canonical alternative (e.g. *montar ??**en** [a] caballo*; 'to ride a horse').

Since the rules used to query collocations were fed with the lemmas of the base and the collocate, these were assigned the frequencies of their respective lemmas also – or of their pattern-matching substitute in the cases where that was necessary. These frequencies were used to calculate the expected frequency for each collocation, which in turn was used to calculate their MI score.

In addition to MI scores, we also used the raw frequency of each collocation instead of using a t-score value, unlike Durrant and Schmitt (2009) or Granger and Bestgen (2014). We discarded t-scores for two reasons. In the first place, in the cited

[4] http://nlp.lsi.upc.edu/freeling/

studies t-score is used as an index of frequency, and reasonably so, given the correlation existing between these two dimensions. However, frequency of co-occurrence is a more direct measure in this respect. Secondly, the two studies cited used a t-score threshold in order to guarantee the collocational nature of their samples. Since we had an independent criterion of collocability (annotators' intuition), no such threshold was needed.

In order to compare the collocations of native-speakers and learners according to their MI-scores and frequency of co-occurrence we grouped them into bands. The bands for MI information were established according to the following system:

(2) MI: band 0 includes values ≥ 0 and < 1; band 1 includes values ≥ 1 and < 2, etc.

As far as frequency of co-occurrence is concerned, given the fact that due to the size of our reference corpus the range of possible values was enormous (from 0 to 205,744 occurrences), we used the natural logarithm of these values in order to obtain a more manageable range sorted according to a non-arbitrary scale. The frequency values thus obtained were arranged in bands as follows:

(3) Frequency of co-occurrence: band 0 includes values ≥ 0 and < 1; band 1 includes values ≥ 1 and < 2, etc.

This arrangement into frequency and MI bands makes it easier to compare native speakers' and learners' data and a similar solution has been adopted by, for instance, Durrant and Schmitt (2009) and Granger and Bestgen (2014). The band size is, however, different in each case.

4. Results

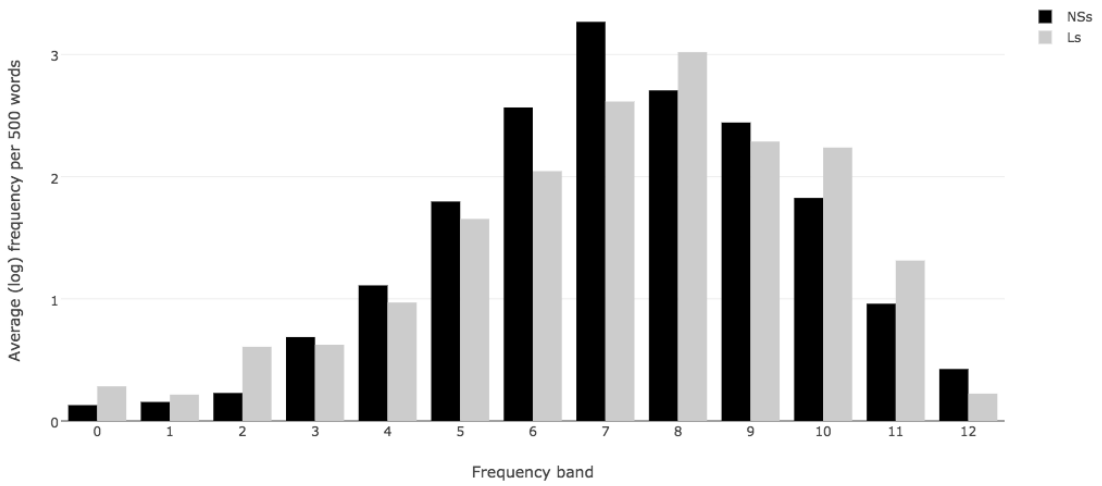
This section presents the results of comparing learners' and native speakers' collocation samples in terms of their respective frequencies and MI scores. As has been done in previous studies, data regarding collocation tokens (i.e., each collocation present in the running text, regardless if it has previously occurred in the corpus or not) was dealt with separately from that regarding collocation lemmas (i.e., one form that groups the different occurrences of the same combination of lexical items, including inflectional

variants).[5] In the case of tokens we studied their distribution in learners' and native speakers' texts, which deviates somewhat from the practice of previous studies (Durrant and Schmitt [2009]; Granger and Bestgen [2014]), as these focus on the proportion of collocations of different frequencies within the total of collocations employed in each text. In the case of collocation lemmas, we focused on their proportion in the repertoire of each subject, much as was done in the case of the cited articles.

4.1 Distribution of tokens

Fig. 1 displays the results of comparing the distribution of collocation tokens belonging to the different frequency bands we distinguish in the native and learner subcorpora. For each text we have calculated the frequency normalised per 500 words of the collocations pertaining to each frequency band and have obtained the mean frequencies of both subcorpora.

Fig. 1. Distribution of collocations in native speakers (NSs) and learners (LSs) according to their frequency



The first three bands correspond to combinations scarcely attested in the reference corpus: it must be kept in mind that a log frequency of 0 corresponds to a frequency of 0, since we have added 1 to all the values in order to transform them into logarithms. A

[5]It should be noted that the cited studies distinguished between tokens and types, but given the partially different nature of the combinations studied here and the morphological complexity of Spanish, it seemed reasonable to focus on tokens and lemmas.

log frequency of 2 corresponds to a frequency of roughly 6. Collocations belonging to these first three bands are more frequent in the learner subcorpus. Band 3 (i.e., log frequency ≥3 and <4) seems to be a turning point, since the collocations in bands 3 to 7 are more frequent in the native subcorpus. The tendency changes again from band 8 onwards, albeit with fluctuations (bands 9 and 12): the collocations belonging to these last frequency bands are used more often by learners (with the previously mentioned exceptions).

The overall tendency, therefore, is for learners to use collocations belonging to the highest frequency bands more often than native speakers, whereas the latter resort more often to collocations of moderate frequencies. As far as scarcely or completely unattested collocations are concerned, they are again more prevalent in the learners' texts.

In order to establish whether these distributional differences were significant from a statistical point of view, we applied the Wilcoxon-Mann-Whitney test. This test has recently been recommended to discover differences in the distribution of words or n-grams within different corpora (Kilgarriff, 2001; Paquot and Bestgen, 2009; Lijffijt et al. 2014): the input for the test are the frequency counts of a given word form or n-gram in the different sections of two or more corpora (either the sections are of equal size or the counts are normalized) transformed into ranks. In our case, instead of a given word form or n-gram, we compared collocations belonging to a given frequency band. We rearranged the initial frequency bands into three groups, the cut-off point being the frequency bands where a change of tendency was observed: collocations of low frequency (L) comprising bands 0 to 2; collocations of moderate frequency (M) comprising bands 4 to 7; and finally collocations of high frequency (H) comprising bands 8 to 12. The results are displayed in Table 2. We adopted the 0.05 significance level corrected for multiple comparisons (0.017, in this case).

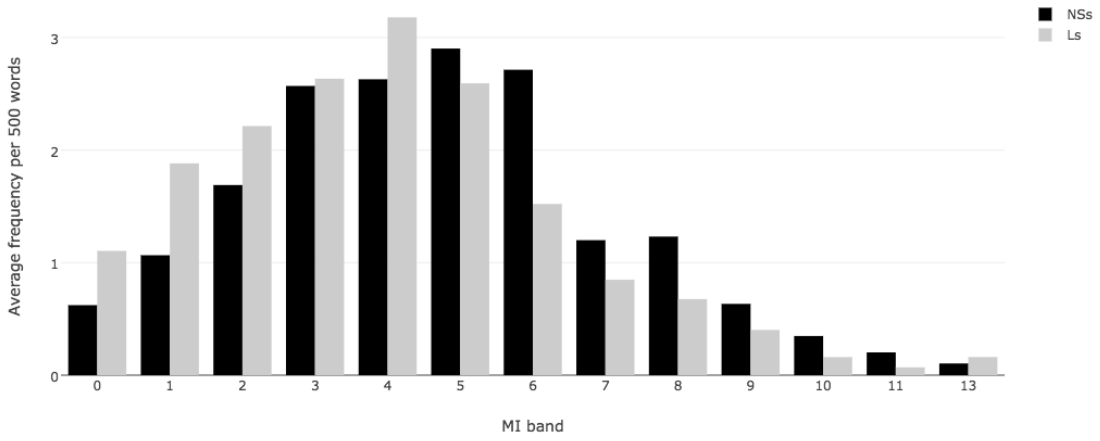
Table 2. Results of the Wilcoxon-Mann-Whitney test of the distribution of collocations with low, moderate and high frequency

Frequency Band	L	M	H
p-value	<0.001	0.4102	0.002

We can see that the central group, that consisting collocations with moderate frequency, seems to be similarly distributed across the texts of both native and learner samples. It is the groups of infrequent and very frequent collocations, both of which are found to a greater extent in learners' texts, which display significantly different distributions across our two samples.

For the comparison of the collocations according to their MI values we followed a similar procedure, but in this case we discarded collocations with less than 5 occurrences in our reference corpus. In addition, the plot reflects only those bands attested on at least 5 occasions in each subcorpus. The discarded bands are mostly cases of isolated extreme values. Thus, for instance, the learner subcorpus contained cases of collocations with MI values under -1, that is, combinations of words that seem to repel each other. Some examples of this category are unusual combinations with more idiomatic alternatives such as *cambiar al cristianismo* lit. 'change to christianity' (MI=-2.08), instead of *convertirse al cristianismo; cambiar a la [verdadera] religión* lit. 'change to the [true] religion' (MI=-2.86), again instead of *convertirse a…*; or *establecer un sufrimiento* 'to set up [cause]? a suffering' instead of *provocar/causar un sufrimiento*. On the other hand, in the native subcorpus there is one collocation with MI above 19 (*colmo de males* lit. 'top of bad things' ≈ 'to top it all'), whereas the collocation with the highest MI in the learner subcorpus is *tomar gazpacho* ('to have gazpacho') with 14.70. The results of this comparison can be seen in Fig. 2.

Fig. 2. Distribution of collocations in NSs and learners according to their MI scores



This time the turning point seems to be the band of MI 5. Collocations with MI values of 5 or higher are more commonly used in the native corpus (with one exception: band 13 with only 10 cases in the learner corpus and 7 in the native one).

In order to ascertain if these two tendencies (i.e. learners' overuse of collocations with MI lower than 5 and their underuse of collocations with MI equal to or above 5) represented a significant difference in the distribution of collocations with different MI scores, we rearranged our sample into two large groups (collocations with MI above and under 5) and again applied the Wilcoxon-Mann-Whitney test to the normalized frequency (per 500 words) of the collocations pertaining to these two groups.[6] The results are shown in Table 3.

Table 3. Results of the Wilcoxon-Mann-Whitney test of the distribution of collocations with MI <5 and MI ≥ 5

MI band		MI <5	MI ≥ 5
p-value		<0.001	<0.001
mean frequency per text	Native Speakers	8.60	9.48
	Learners	11.38	6.29

The distribution of the two groups of collocations differs significantly between the two samples. In this case, we also added the (normalized) mean frequency in the texts of learners and native speakers, since the sample is slightly different to that plotted in Fig. 2. These data not only show that collocations with an MI lower than 5 have a higher mean frequency in learners' texts, but also that they are used more often than collocations with an MI of 5 or over.

4.2. Repertoire of collocational lemmas in learners and native speakers

Like Durrant and Schmitt (2008) or Granger and Bestgen (2014), we also examined the collocational repertoire of learners and native speakers in terms of their composition by type – or perhaps, more precisely in our case, by collocation lemma. In this case, instead

[6] This time all the collocations with a frequency equal to or higher than 5 in the reference corpus were included.

of looking at their distribution in the texts of the two subcorpora, we analysed the proportion of each band out of the total of collocations in each sample. The results are not markedly different from the previous ones, since most of the collocations only occurred once (repetitions were more common in the learner subcorpus, which is also made up of longer texts). After excluding the repetitions of collocations with the same lemmas for base and collocate, the resulting samples contained 1,000 lemmas in the case of native speakers and 1,529 in the case of learners. To examine the proportion of lemmas with different MI values the samples were further reduced, as collocations with less than 5 occurrences in our reference corpus were again discarded, yielding a sample of 986 and 1,490 lemmas for native speakers and learners respectively.[7]

Fig. 3 displays the mean proportions per text of the collocation lemmas of each frequency band whilst Fig. 4 shows the same information, but this time with the collocation lemmas arranged according to their MI values in our reference corpus.

Fig. 3. Percentage of collocation lemmas per text according to their frequency

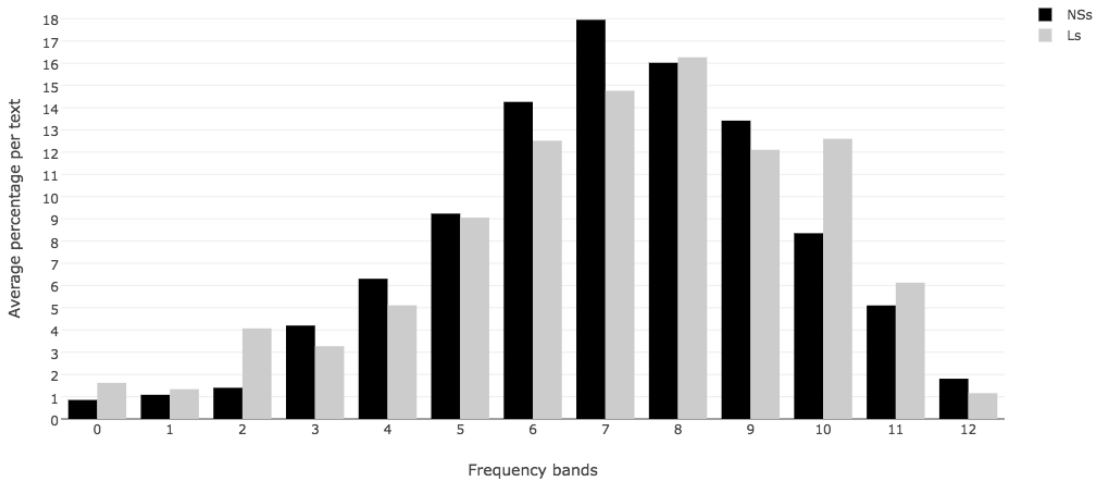
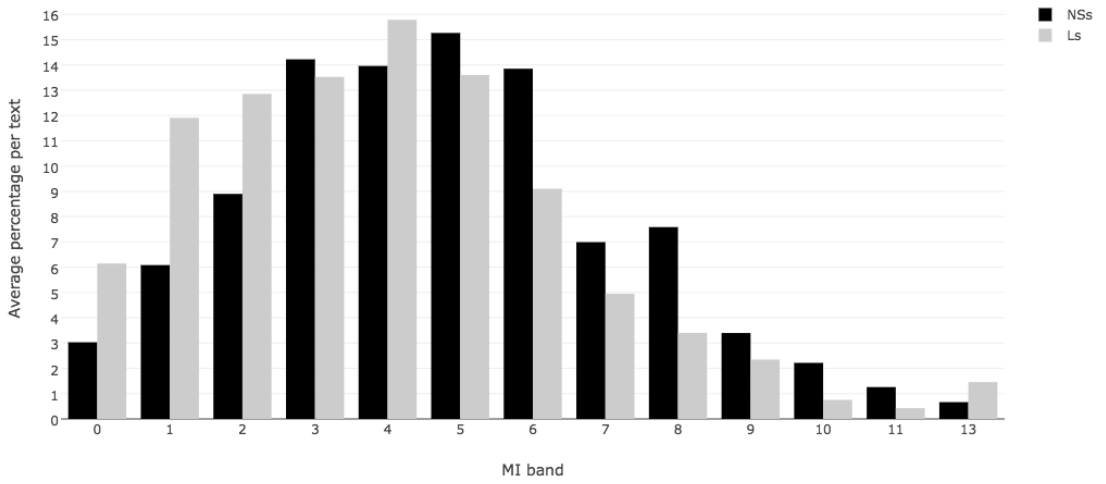


Fig. 4. Percentage of collocation lemmas per text according to their MI band

[7] It should be noted that we only eliminated repetitions of the same lemma in the same text, not in the whole corpus.



As mentioned above, the tendencies are very similar to those discovered when examining the distribution of tokens. Lemmas with very low frequencies (bands 0 to 2) are more frequent in learners' texts, whilst lemmas with moderate frequencies (bands 3 to 7) are more common in native speakers. Finally, collocation lemmas with the highest frequencies (from band 8 onwards) seem to make up a larger proportion of the learners' repertoire, although with exceptions (bands 9 and 12).

As far as MI is concerned, collocation lemmas with MI values equal to or higher than 5 are consistently more used by native speakers. Lemmas with MI lower than 5 are in general more common in learner texts with one exception (band 3).

We proceeded as before to test the significance of the observed differences. In the case of frequency, we regrouped the initial bands into the three groups that showed different tendencies in learners and native speakers, i.e. low frequency (bands 0 to 2), moderate frequency (bands 3 to 7) and high frequency (bands 8 to 12) and we applied the Wilcoxon-Mann-Whitney test (alpha=0.017) to the proportions of collocations registered in each text of the two subcorpora and belonging to each of these three groups. On this occasion, the differences between collocations of low and moderate frequency turned out to be significant ($p=0.001$ and $p=0.005$ respectively) not so the difference observed in the collocations belonging to the high frequency group.

As for MI, we again grouped the collocations into two sets – those with MI < 5, and those with MI ≥ 5 – and tested the differences between them using the same test. In both cases the differences are significant ($p<0.001$). Again, the collocations with MI < 5 seem to be overused by learners (the mean percentage per text is 46.37%, as opposed to

61.08% in the case of native speakers), whilst the collocations with MI ≥ 5 occur less frequently in their texts (with a mean percentage of 36.50%, as opposed to 52.13% in the native subcorpus).[8]

5. Discussion

In terms of their frequency in the reference corpus, both the unattested or hardly attested collocation tokens and lemmas were more often found in learner texts. It is not surprising that a large proportion of the unattested collocations are instances of non-native-like combinations. Thus, most of the unattested combinations were deemed as erroneous by the annotators of the corpus: out of the 28 collocations belonging to the 0 band, 23 (ca. 82%) were considered incorrect. The proportion of incorrect collocations pertaining to the lowest three frequency bands (0, 1 and 2; i.e., the ones that presented greater frequencies in the learner corpus both in terms of lemmas and tokens) is significantly larger than the proportion of incorrect collocations from the rest of frequency bands: 65 out of 107 as opposed to 331 out of 1592; or, in percentages, 60.7% vs. 20.8% respectively.[9]

Some examples of the incorrect collocations belonging to these extremely infrequent groups are unidiomatic combinations such as *gritar abusos* lit. 'to shout abuses', instead of *lanzar insultos* lit. 'to throw insults', *extender una colección* lit. 'to extend a collection', instead of *ampliar una colección*, etc., or simply combinations made up of non-words, such as *las temperaturas *frescan* instead of *las temperaturas refrescan* 'the temperature falls down', lit. 'freshens', or *capitón de la equipe* instead of *capitán del equipo* 'team leader'.

As noted, not all the infrequent collocations were considered incorrect by the annotators: learners managed to produce natural sounding yet unattested combinations such as *acertijo obvio* 'obvious riddle' or *acabar el PhD* 'to finish the PhD', the latter including an English denomination seemingly more and more usual in Spanish.

Rare collocations (bands 0, 1 and 2 again) amounted to a smaller proportion in the native sample (ca. 3%, as opposed to the 6% of learners). In spite of their rarity, they were considered collocations by annotators due to the fact that they responded to the conception of the ECL framework. Thus, even though there are no instances of

[8] It must be remembered that the percentages are calculated over the total of collocations of each text, but the collocations with less than 5 occurrences in our reference corpus were disregarded for the purpose of MI comparison.

[9] A z-test for two proportions yields a z-score of -9.46 and a p-value of 0.000.

transatlántico(s) naufragado(s) 'wrecked liner(s)' in our reference corpus, it qualifies as a collocation, since *naufragar* is a predicate that includes *transatlántico* in its meaning as an exemplar of the class 'ships' (see Mel'cuk 1995: 182). Likewise, *fajo* in *fajo de piastras* 'bundle of piasters' is a quantifying noun that typically combines with other nouns meaning 'bank notes'. The fact that the piaster is an uncommon currency unit in the life of Spanish speakers probably account for the absence of the collocation at issue in our corpus.

The results for infrequent collocations are in contrast to those obtained in Durrant and Schmitt (2009) and Granger and Bestgen (2014). Both of these studies find that unusual combinations are more often used either in texts written by native speakers or learners with a more advanced proficiency level. It must be kept in mind, however, that the three studies are not strictly comparable. Both Durrant and Schmitt (2009) and Granger and Bestgen (2014) focus on adjacent word pairs, i.e., bigrams: pre-modifiers and nouns in the case of the former, and all the bigrams of their texts in the case of the latter (with particular attention to certain categories; see above). Furthermore, we have limited ourselves to combinations that previously had been manually identified as phraseological collocations.

In contrast to the significant results for the groups of both low frequency tokens and lemmas, the differences between the groups of intermediate frequency were only significant in the lemma comparison whilst the differences between the high frequency groups were only significant in the token comparison. This could be suggestive of the fact that learners make use of a similar repertoire of high frequency collocations more often than native speakers; in other words, they stick to high frequency combinations and repeat them more often than native speakers. This is in line with the idea that frequent combinations are some kind of lexical "teddy bears" for learners (Nesselhauf, 2005: 69, borrowing the term from Hasselgren, 1994, who used it for single-word vocabulary). However, one must be cautious in this respect for two reasons. The first concerns the different procedures we followed in the comparisons of lemmas and tokens: in the case of the latter, we compared their distribution in the texts of learners and native speakers; as for the former, on the other hand, what was compared was the proportion within the whole repertoire of collocations of both groups. Secondly, and perhaps more importantly, there is the fact that native and learner texts differ in length, the latter being longer on average, so that learner texts provide more occasions for repetition.

By way of a brief recapitulation of what we have seen so far, it can be said that the writing of learners of Spanish is characterised in the first place by the presence of a higher number of collocations that are unattested in our reference corpus, frequently showing deviant features. Their repertoire of collocations of moderate frequency is narrower than that of native speakers, but they make the most of them, to the extent that no significant difference can be detected in terms of token distribution. Lastly, very frequent combinations constitute an equivalent proportion of the collocational repertoire of the two populations, but learners overuse them in a significant way. The results concerning the overuse of high frequency collocations by leaners are in line with previous studies (Bestgen and Granger, 2014; Durrant and Schmitt, 2009; Granger and Bestgen, 2014; Vincze et al., in press).

As regards the comparison based on the MI scores of the collocations of learners and native speakers, similar results were obtained for both token distribution and lemma composition: learners show a smaller repertoire of collocations with MI ≥ 5 than native speakers and these collocations occur less frequently in their texts. Extremely low values of MI (-2, -3) that are attested in the learner sample and absent from the native speaker texts are the like of *quedar en un alojamiento* lit. 'to stay in an accommodation' or *puente de clase* lit. 'bridge of class', meaning 'long weekend', which also happen to be very infrequent.[10] These, however, are isolated cases. Not so rare are combinations with MI between 0 and 3, the conventional threshold of collocability posited for MI. On the contrary, they are quite frequent both in the learner and in the native sample (29.4% and 17.6% of the total, respectively). Some of the collocations most frequently repeated by learners belong to this group: e.g., *tener padre* 'to have a father'; *tener año(s)* lit. 'to have year(s)', 'be X year(s) old'); *tener culpa* lit. 'to have [the] blame', 'to be one's fault', amongst others They seem to be frequent combinations in general, made up of equally frequent words and, even if they fall below the threshold of collocability attending exclusively to MI, they are well above the threshold of significance for t-score: *tener padre* has a t-score of 96.05, *tener culpa* of 125.60 and *tener año(s)* of 125.75. Also from the phraseological perspective they qualify as *bona fide* collocations:

[10] The former was considered erroneous, but simply because of a grammatical feature: the annotators proposed the correction *quedar**se** en un alojamiento,* with a reflexive clitic. Otherwise the combination was deemed acceptable, in spite of its low frequency. *Puente de clase* was considered erroneous because *puente* 'long weekend' is normally used without specifications of that nature (*de clase*).

they are combinations of either relational[11] or predicate nouns and support verbs; the meaning of the verb depends on the noun (with kinship nouns, it denotes some sort relation, rather than 'possession'; with a noun such as *culpa* 'blame' it becomes some sort of passive verb, etc.); from a contrastive perspective, some of them can be viewed as idiosyncratic combinations (Sp. *tener años* → Eng. *to be X years old;* Sp. *tener la culpa* → *to be one's fault*), etc.

Collocations with high MI values, by contrast, are not necessarily very frequent *per se*, but they are much more frequent than their constituents would have us expect. Thus, in a pair of synonyms, such as *ir a la escuela* and *asistir a la escuela* (both meaning 'to attend school'), the former has a much lower MI score (2.30 vs 5.20) due to the fact that *ir* is nearly 20 times as frequent as *asistir* and, therefore, it yields a much higher expected frequency for the whole combination than the latter. In this respect, higher MI scores could perhaps be interpreted as evidence of greater coherence as Ellis et al. (2008: 380) claim, but in our case not in terms of grammatical well-formedness or distinctive function, given the more restricted nature of our sample.[12] In the examples at hand, *asistir* is a less polysemous verb than *ir,* with fewer possibilities regarding the contexts in which it can occur and, probably, more limited with respect to the registers in which it can be used (it seems unlikely in informal registers). In terms of meaning *ir* and *asistir* are equivalent in the context of *a la escuela,* since the former not only expresses motion, but 'motion+attending'. If this pair is representative of what differences in MI reveal, such differences should be interpreted in terms of choice of forms made for the expression of a given meaning. High MI scores, then, would point to combinations of forms with less semantic versatility than those occurring in combinations with lower MI scores or simply to combinations made up of unusual forms. A classic example of this type of combination is *miedo cerval* ('great fear'): *cerval* is an adjective that only occurs in the context of a couple of nouns (*miedo, espino* 'hawthorn') and, furthermore, probably limited to literary texts.

All in all, the results of the MI comparison are in line with previous studies (Durrant and Schmitt, 2009; Lorenz, 1999): learners underuse high MI collocations. There are a number of possible explanations for this fact. First, as we have seen, high

[11] Within the MTT framework, relational nouns are considered quasi-predicates since they have "actant" slots: for instance 'father of X', 'age of X' (see Polguère, 2012).
[12] Ellis et al. (2008) work on a sample of n-grams, which can include grammatically incomplete structures or chains crossing grammatical boundaries. Our sample of collocations consists by definition of combinations with a given syntactic structure and with a particular semantic bond.

MI collocations are not necessarily very frequent and, as previous studies (Durrant and Schmitt, 2009; Granger and Bestgen, 2014) have also pointed out, non-native speakers rely heavily on highly frequent combinations. Second, the frequency of the components of the collocations taken separately also seems important. Highly frequent collocations with low MI values – prevailing in learner discourse – are made up of similarly very frequent portmanteau words, such as *tener*, which are likely to have a greater presence in learners' input and thus be easier to acquire. Finally, the constituents of collocations with high MI scores – or at least one of them – are not only quite specific with respect to their meaning, but also in the lexical contexts in which they occur.

6. Conclusion and implications

The results of this study suggest that association measures (and in particular those used here: frequency of co-occurrence and mutual information) are useful to discover patterns in the writing of learners of Spanish that characterise and distinguish them from native speakers. In spite of the different approaches followed and the different target languages, our results are in line with those of previous studies (Lorenz, 1999; Durrant and Schmitt, 2009) focusing on English: as compared to native speakers, learners overuse high frequency collocations but underuse collocations with high mutual information scores. Accordingly, one could hypothesize that frequency has a facilitating effect for the acquisition of collocations by learners, whilst high MI does not seem to be salient for this group of speakers, at least to the extent of enabling the acquisition of collocations characterised by this trait (Ellis et al. 2008 provide psycholinguistic evidence pointing in this direction). Elaborating on this argument, one could predict that high-frequency collocations would be acquired faster by learners and used earlier than collocations with moderate or low frequencies of occurrence, even if they are strongly associated (i.e., they have high MI scores). This is precisely what Granger and Bestgen (2014) found in their pseudo-longitudinal study of learners of English. In this respect, it could be argued that one of the limitations of the present study is its cross-sectional nature and that the confirmation of this hypothesis would require (pseudo-)longitudinal studies of the collocation production of learners of Spanish, which opens up a line for future research.

Our findings have implications for vocabulary teaching in Spanish. If it is true that the difficulty collocations pose to learners is co-related with their frequency and their MI, these two measures present themselves as criteria that should be weighed

when deciding the inclusion and sequencing of collocations in such materials as syllabi, learner dictionaries, textbooks, etc. (the role of frequency in this respect has already been highlighted by Nation, 2001: 329 or Martinez, 2013). So far, the most prestigious guideline in Spanish that attempts to present vocabulary, including collocations, following the indications of CEFR (Council of Europe, 2001), but with a greater degree of specification, is the Instituto Cervantes's (1997-2016) *Plan Curricular*. It bases the sequencing with which vocabulary is presented on the intuition that experienced teachers have about its frequency and usefulness. Collocation lists extracted from corpora and arranged by means of association measures could provide Spanish teachers with larger and more systematic repertoires of collocations, as well as with criteria for their importance and grading. In this respect, Ferrando Aramo (2012: 360) regrets a lack of studies of collocational frequency in Spanish and lists based thereon.

Further applications could include automated correction or scoring (see Granger and Bestgen, 2014: 248). Within the field of automated correction the use of association scores, even in Spanish, already seems to be well established. The tool *HaRenEs,* for instance, bases its corrections exclusively on association measures obtained from frequency data (Ferraro et al., 2014). As far as automated scoring is concerned, if a correlation is proven between the proportion of very frequent collocations, the proportion of high MI collocations and learners' proficiency level (as in Bestgen and Granger, 2014), these two parameters could be useful in determining the proficiency level of test candidates, but as stated before, there is still a need for (pseudo-)longitudinal studies to be carried out in this particular field.

References:

Alonso Ramos, Margarita, Leo Wanner, Nancy Vázquez Veiga, Orsolya Vincze, Estela Mosqueira Suárez & Sabela Prieto González. 2010a. Tagging collocations for learners. In Sylviane Granger & Magali Paquot (eds.), *Elexicography in the 21st Century: New Challenges, New Applications. Proceedings of eLex2009* (Cahiers du Cental 7), 375–380. Louvain-la Neuve: Presses Universitaires de Louvain.

Alonso Ramos, Margarita, Leo Wanner, Orsolya Vincze, Gerard Casamayor, Nancy Vázquez Veiga, Estela Mosqueira Suárez & Sabela Prieto González. 2010b. Towards a motivated annotation schema of collocation errors in learner corpora. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the*

Seventh Conference on International Language Resources and Evaluation (LREC'10), 19-21 May 2010, Valletta, Malta, 3209–3214. European Language Resources Association (ELRA).

Altenberg, Bengt & Sylviane Granger. 2001. The Grammatical and Lexical Patterning of MAKE in Native and Non-Native Student Writing. *Applied Linguistics* 22, 173–195.

Bestgen, Yves & Sylviane Granger. 2014. Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing* 26, 28–41.

Cowie, A. P. 1981. The treatment of collocations and idioms in Learners' dictionaries. *Applied Linguistics* 2(3), 223-235.

Cowie, A. P. 1998. Phraseological dictionaries: Some East–West comparisons. In A. P. Cowie (ed.*), Phraseology. Theory, analysis, and applications*, 209–228. Oxford: Oxford University Press.

Durrant, Philip. 2008. *High frequency collocations and second language learning*. Nottingham: University of Nottingham dissertation.

Durrant, Philip. 2014. Corpus frequency and second language learners' knowledge of collocations. *International Journal of Corpus Linguistics* 19(4), 443-477.

Durrant, Philip & Norbert Schmitt. 2009. To what extent do native and non-native speakers maker use of collocations? *IRAL* 47, 157–177.

Ellis, Nick, Rita Simpson-Vlach & Carson Maynard. 2008. Formulaic Language in Native and Second Language speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *TESOL Quarterly* 42(3), 375–396.

Erman, Britt and Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text* 20(1), 29–62.

Evert, Stefan. 2005. *The Statistics of word cooccurrences*. Stuttgart: Universität Stuttgart dissertation.

Evert, Stefan. 2008. Corpora and collocations. In Anke Lüdeling y Merja Kytö (eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter http://purl.org/stefan.evert/PUB/Evert2007HSK_extended_manuscript.pdf (accessed 15 April 2016)

Ferrando Aramo, Verónica. 2012. *Aspectos teóricos y metodológicos para la compilación de un diccionario combinatorio destinado a estudiantes de E/LE*. Tesis doctoral. Tarragona: Universitat Rovira i Virgili dissertation.

Ferraro, Gabriela, Rogelio Nazar, Margarita Alonso Ramos & Leo Wanner. 2014. Towards advanced collocation error correction in Spanish learner corpora. *Language Resources and Evaluation* 48, 45–64.

Firth, John R. 1957 [1951]: "Modes of Meaning". In John R. Firth, *Papers in Linguistics 1934-1951,* 190–216. London: Oxford University Press.

Gilquin, Gaëtanelle. 2007. To err is not all: What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift Für Anglistik Und Amerikanistik 55*(3), 273–291.

Granger, Sylviane & Yves Bestgen. 2014. The use of collocations by intermediate vs. advanced non-native writers: A bi-gram based study. *IRAL* 52(3), 229-252.

Granger, Sylviane & Magali Paquot. 2012. Formulaic Language in Learner Corpora, *Annual Review of Applied Linguistics* 32, 130-149.

Halliday, M.A.K. 1966. Lexis as a linguistic level. In C.E. Bazell, J.C. Catford, M.A.K. Halliday & R.H. Robins (eds.), *In Memory of J.R. Firth*. London: Longman, 148–162.

Hasselgren, Angela. 1994. Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 4 (2), 237–260.

Hausmann, Franz Josef. 1989. Le dictionnaire de collocations. In Franz Josef Hausmann, Oskar Reichmann, Hans Ernst Wiegand, & Ladislav Zgusta (eds.), *Wörterbücher : ein internationales Handbuch zur Lexikographie. Dictionaries. Dictionnaires*, 1010–1019. Berlin: Mouton De Gruyter.

Kilgarriff, Adam & Irene Renau. 2013. *esTenTen, a Vast Web Corpus of Peninsular and American Spanish. Procedia 95,* 12–19.

Kilgarriff, Adam. 2001. Comparing Corpora. *International Journal of Corpus Linguistics* 6(1), 97–133.

Krishnamurthy, Ramesh. 2000. Collocation: from *silly ass* to lexical sets. In Chris Heffer and Helen Sauntson (eds.), *Words in context: A tribute to John Sinclair on his retirement,* 31–47. Birmingham: University of Birmingham.

Lexical Computing. 2015. Statistics used in the Sketch-Engine. https://www.sketchengine.co.uk/wp-content/uploads/ske-stat.pdf, (accessed 1 June 2016)

Lijffijt, Jefrey, Terttu Nevalainen, Panagiotis Papapetrou, Kai Puolamaki & Heikki Mannila. 2014. Significance testing of word frequencies in corpora. *Digital*

Scholarship in the Humanities, fqu064. http://dsh.oxfordjournals.org/content/early/2014/12/08/llc.fqu064 (accessed 4 December 2015)

Lorenz, Günter. 1999. *Adjective Intensification - Learners versus Native Speakers: A Corpus Study of Argumentative Writing*. Amsterdam: Rodopi.

Lozano, Cristobal. 2009. CEDEL2: Corpus Escrito del Español L2. In Carmen M. Bretones Callejas, José Francisco Fernández Sánchez, José Ramón Ibáñez Ibáñez, Maria Elena García Sánchez, Mª Enriqueta Cortés de los Ríos, Sagrario Salaberri Ramiro, Mª Soledad Cruz Martínez, Nobel Pedrú Honeyman, & Blasina Cantizano Márquez (eds.), *Applied Linguistics Now: Understanding Language and Mind / La lingüística aplicada hoy: Comprendiendo el lenguaje y la mente*, 197–212. Almería: Universidad de Almería.

Lozano, Cristobal, & Mendikoetxea, Amaya. 2013. Learner corpora and second language acquisition: The design and collection of CEDEL. In Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson (eds.), *Automatic Treatment and Analysis of Learner Corpus Data*, 65–100. Amsterdam: John Benjamins.

Martinez, Ron. 2013. A framework for the inclusion of multi-word expressions in ELT. *ELT Journal 67*: 184–198.

Mel'čuk, I. A. (1995). Phrasemes in language and phraseology in linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms: Structural and Psychological perspectives* (pp. 167–232). Hillsdale, NJ: Lawrence Erlbaum.

Mel'cuk, Igor. 2012. Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology* 3(1): 31–56.

Nation, Paul. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nesselhauf, Nadia. 2005. *Collocations in a Learner Corpus*. Amsterdam/Philadelphia: John Benjamins.

Paquot, Magali & Yves Bestgen. 2009. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In Andreas H. Jucker, Daniel Schreier and Marianne Hundt (eds.) Corpora: Pragmatics and Discourse, 243-265. Amsterdam: Rodopi.

Pawley, Andrew. 1985. On speech formulas and linguistic competence. *Lenguas Modernas*, *12*, 84–104.

Polguère, Alain. 2012. Propriétés sémantiques et combinatoires des quasi-prédicats sémantiques. *Scolia 26*, 131–152.

Sinclair, John M. (ed.). 1987. *Looking Up - An account of the COBUILD Project in lexical computing*. London: Harper-Collins

Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2(1), 23–55.

Vincze, Orsolya. 2015. *Learning multiword expressions from corpora and dictionaries*. A Coruña: Universidade da Coruña dissertation.

Vincze, Orsolya & Margarita Alonso Ramos. 2013. Incorporating frequency information in a collocation dictionary: Establishing a methodology. *Procedia - Social and Behavioral Sciences* 96, 241–248.

Vincze, Orsolya, Margarita Alonso Ramos & Estela Mosqueira. 2011. Exploiting a learner corpus for the development of a CALL environment for learning Spanish collocations. In Iztok Kosem & Karmen Kosem (eds.), *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011,* 280–285. Ljubljana: Trojina, Institute for Applied Slovene Studies,.

Vinzce, Orsolya, Marcos García Salido, Ana Orol & Margarita Alonso Ramos. In press. A corpus study of Spanish as a Foreign Language learners' collocation production. In Margarita Alonso Ramos (ed.), *Spanish Learner Corpus Research: Current trends and Future Perspectives*. Amsterdam/Philadelphia: John Benjamins.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*, Cambridge: Cambridge University Press.