Towards a Graded Dictionary of Spanish Collocations

Marcos García Salido, Marcos Garcia, Margarita Alonso-Ramos

Universidade da Coruña, CITIC, Grupo LyS, Dpto. de Letras,
Fac. de Filoloxía. 15071, A Coruña
E-mail: {marcos.garcias, marcos.garcia.gonzalez, margarita.alonso}@udc.gal

Abstract

Several recent studies have observed that texts of different quality and written by learners at different proficiency levels also vary in the lexical combinations they contain. Such variation can be operationalized by quantitatively measuring the association between the components of these lexical combinations. In particular, pointwise mutual information (MI) has proved to be a good predictor of proficiency development, as several studies on English learners' writing have shown. This paper examines whether association measures are also a good predictor for the proficiency level of texts written by learners of Spanish, with a view to using such information for grading lexical combinations in order to include them in a collocation dictionary of Spanish. The study also investigates whether the association measures that correlate with learners' proficiency level can discriminate between phraseological collocations and non-collocations. Our results show that, whereas the MI of learner texts' lexical combinations is a better predictor of author proficiency than frequency, the latter performs better in identifying phraseological collocations among the whole set of lexical combinations.

Keywords: graded collocation dictionary; CEFR proficiency level; association measures

1. Introduction

Phraseological expressions permeate discourse to a considerable extent. Erman and Warren (2000) estimate that, on average, 55% of texts is made up of prefabricated expressions. Collocations surely are a subset among those prefabricated expressions and are therefore an essential component of learning a new language. In fact, several studies have found that collocations are a challenging aspect of language learning: see Granger (1998), Nesselhauf (2004), or Vincze et al. (2016), to cite but a few.

Despite the importance of collocations in language learning, the attention given to this phenomenon in curricula or assessment materials is not always evident, as noticed by Paquot (2018). According to her, the Common European Framework of Reference for Languages—henceforth CEFR (Council of Europe, 2001)—assumes a very traditional understanding of phraseology, by obviating frequent word combinations and using the term *phrase* mostly for stock phrases and pragmatically conditioned expressions. Paquot emphasises that, by ignoring learners' phraseological competence, we are losing a valuable assessment criterion for language proficiency.

In the particular case of Spanish, Higueras García (2017) argues that, whereas research has devoted considerable attention to collocations, these combinations are still not very well treated in Spanish Language Teaching. She also favours a flexible conception of the notion, which includes frequently used combinations, even if they are not properly the result of lexical restrictions. As for their introduction to learners, although she mentions frequency, Higueras clearly prefers other selection criteria, such as their relation with syllabus' topics and with communicative functions.

Even though, from the situation Higueras García (2017) depicts, collocations seem to be a phenomenon rather neglected by the Spanish Teaching community in general, learners of this language have some reference works at their disposal. The *Diccionario de colocaciones del español* (DiCE; Alonso-Ramos, 2004) is an online dictionary that follows the principles of the Meaning-Text Theory in the treatment of collocations. It is an ongoing project that so far incorporates lexical units related to the sentiment's lexical field and is in the process of including academic collocations. The sentiment collocations provide users with CEFR level indications. The *Diccionario combinatorio práctico del español contemporáneo* (PRÁCTICO; Bosque, 2006) is a paper dictionary based on a mostly theoretical combinatorial dictionary (REDES, Bosque, 2004). In its structure, it is more similar to other learner-targeted collocation dictionaries, such as Benson et al. (1986), than REDES. In contrast to DiCE, and in spite of being corpus-based, PRÁCTICO in general does not provide notes on collocation frequency, but occasionally indicates their semantic prosody. Finally, the *Herramienta de Ayuda a la Redacción en español* (HARenEs; Alonso Ramos et al., 2015) is a web tool that gives its users collocations directly extracted from a corpus in a more user-friendly manner than concordancers.

This paper explores the possibilities of lexical association measures in grading learners' lexical combinations and in identifying phraseological collocations among such combinations. Its final aim is to explore a method to compile a collocation dictionary that combines features offered separately by some of the reference works reviewed: firstly, by including a vast set of collocations representative of Spanish, like PRÁCTICO; and secondly, by offering notes on the CEFR level of collocations, like DiCE, providing thus guidelines to the Spanish teaching community for grading lexical contents and for assessment. In what follows, we review some related work in Section 2. Next, the method proposed is described in Section 3. Section 4 presents and discusses the results and evaluates the viability of the method for compiling a graded collocation dictionary of Spanish before moving onto the conclusions (Section 5).

2. Lexical combinations, frequency-based association measures and proficiency

When it comes to grading vocabulary, lexical frequency shows up repeatedly as a useful criterion (Nation, 2001; Alvar Ezquerra, 2005). The rationale behind this recommendation is that the most frequent vocabulary of a language covers larger proportions of text than less frequent units (be they word-families, lemmas or word-forms). Consequently, learning this first would theoretically lead to great advances in understanding and producing texts. Frequency has been proposed as a grading criterion for multi-word vocabulary as well. Martinez (2013) suggests to give priority to lexical combinations that are both frequent and semantically opaque.

Frequency as a grading criterion has been applied to several vocabulary repertoires directed to Spanish teaching. The *Plan curricular del Instituto Cervantes* (henceforth, PCIC)[1] is a set of guidelines that adapts the recommendations of the CEFR with a greater degree of specificity. Several of its sections provide vocabulary graded by proficiency—including some collocations. This document states that vocabulary selection is based on frequency and usability as perceived by experienced professionals, among other criteria. In a similar vein, corpus frequency has been used for grading the collocations included in at least two collocation dictionaries: the DiCE (García Salido & Alonso Ramos, 2017) and the *Dizionario delle Collocazioni Italiane per Apprendenti* (Spina, 2016)—in the case of the latter, frequency has been used along with aspects such as the topic to which the vocabulary is related.[2]

Irrespective of whether most frequent lexical combinations are taught first, examination of learners' production seems to back the idea that such combinations are acquired and used earlier than less frequent ones. Thus, in an analysis of texts of intermediate and advanced learners of English, Granger and Bestgen (2014) observe that the first group uses a larger proportion of bi-grams with high t-score values[3] than the latter. However, that is not the whole story. Granger and Bestgen (2014) also analyse their learners' bi-grams in terms of another association measure: pointwise mutual information (MI), which results from the ratio between the observed and expected frequencies of a combination. In this case, it is advanced learners who use the combinations with highest MI values more often. In a further study, Bestgen and Granger (2014) fail to observe a significant correlation between t-score and text quality as determined by professional English teachers, but they do find a significant positive correlation between quality and MI.

More recently, Paquot (2018) has studied the phraseological use of advanced learners of English (levels B2 through C2) and found that the MI of lexical combinations used in learner texts predicts teachers' ratings better than any other measure of syntactic or lexical complexity. In contrast to the earlier references, in this study Paquot focuses on combinations of two lexical units related by a syntactic dependency (namely, verb plus

[1] https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/

[2] Here, we limit ourselves to review collocation dictionaries somehow including CEFR level information, since offering a complete picture of vocabulary selection and grading, as undertaken in projects such as the English Vocabulary Profile (Capel, 2010, 2012), falls outside the scope of this paper.

[3] A measure highly correlated with co-occurrence frequency.

object noun, verb plus modifying adverb and noun plus modifying adjective), rather than on bigrams.

In summary, whereas priming frequent phraseological combinations and introducing them first in curricula seems reasonable by virtue of its usability, MI seems a better predictor of proficiency when examining learner production.

3. Methodology

In what follows we describe the methods used to explore to what extent association measures (AMs) predict the proficiency of learners of Spanish in order to apply this information in the compilation of a graded collocation dictionary. With this aim we extracted lexical combinations from a corpus of learner texts and assigned them the AMs corresponding to those very combinations in a reference corpus of Spanish.

The learner corpus used in this study comes from CEDEL2 (Lozano & Mendikoetxea, 2013). We chose texts whose authors got a score of 50% or higher in a placement test[4] administered to them at the time the corpus was compiled. Also, we included only texts that had a length of at least 200 words. The resulting sample consisted of 234 texts comprising 102,621 words. These were graded according to CEFR levels by three expert teachers of Spanish as a Second/Foreign Language who reached a consensus of 67% (Krippendorff's alpha = 0.7). Texts were assigned the level chosen by the majority of raters. The distribution of texts across levels was unequal, as can be seen in Table 1.

CEFR level	A1	A2	B1	B2	C1	C2
No. of texts	23	66	95	39	8	4

Table 1: No. of texts by CEFR level.

There is a clear relationship between authors' median scores in the placement test and the grading of their texts by experts, with the exception of levels C1 and C2, where this correspondence is reversed, as can be seen in Figure 1. On the other hand there is a considerable overlap between the scores obtained by authors of B2 through C2 texts, on one hand, and A1 and A2 texts, on the other. Likewise, B1 texts correspond to a spread range of scores in the placement test.

This learner corpus was tokenized and lemmatized by means of LinguaKit (Garcia & Gamallo, 2015) and PoS-tagged using FreeLing (Padró & Stanilovsky, 2012). Lemmas and PoS-tags were manually revised in order to assign existent Spanish forms to possible lemmas. Only those forms that could be identified with a Spanish word beyond reasonable doubt were corrected.[5] These data were then submitted to syntactic parsing

[4] The test in question is a standardized level-placement test developed by the University of Wisconsin (Lozano & Mendikoetxea, 2013).

[5] For instance, the token *siguente* was lemmatized as the canonical *siguiente* 'next', but

using UDPipe models (Straka et al., 2016).

From this corpus we extracted pairs of lemmas in the following syntactic dependencies: object-verb (obj), subject-verb (nsubj) and noun plus modifying adjective (amod). The collocation candidates extracted from the learner corpus were then assigned the association measures corresponding to these very collocations in a reference corpus. The reference corpus from which the measures were extracted was a 170-million word fragment of Mark Davies' *Corpus del español*[6] automatically processed with the same tools used for the learner corpus—but without manual supervision this time. In spite of the variety of lexical association measures available, we chose MI and frequency given their previous use as predictors of proficiency level in several studies, namely Bestgen and Granger (2014), Granger and Bestgen (2014), and Paquot (2018),[7] as noted in Section 2, even though some other measures might perform better in the detection of phraseological combinations (Pecina, 2010).

These data were then used to fit a generalized linear mixed model. The association measures of the combinations that reached a frequency of 3 or higher[8] in the reference corpus were the independent variables of the model, and the dependent variable was the CEFR levels assigned by the teachers to the texts where they appeared. For this analysis we tried different solutions, namely: (i) using the AMs of the whole set of combinations of each text meeting the conditions already mentioned; (ii) assigning a mean score to each text based on the combinations of each dependency type; and (iii) calculating a unique mean score based on the three dependencies considered taken together. Only in the latter case did we obtain significant results (see Section 4 below).

Additionally, all the lexical combinations extracted from the learner corpus and in the above-mentioned syntactic dependencies (plus noun–preposition–noun) were manually revised in order to identify those that qualified as phraseological collocations. For this, the annotator followed Meaning-Text Theory's definition of *collocation* (Mel'čuk, 2012). According to it, collocations are compositional phrasemes consisting of two elements: one freely chosen by speakers (the base); the other (collocate), which predicates some meaning of the base, is selected depending on the latter: cf. Sp. *vuelta* and its English equivalent *walk*, which cannot be combined with the direct translations of their respective support verbs: *dar una vuelta* 'lit. *give a walk' vs. **tomar una vuelta* 'lit. take a walk', in spite of the sense equivalence of the two expressions. This definition

contesto used as a noun was left as was, since it seems a clear calque from English quite removed from its Spanish equivalente *concurso*, and only its tag was changed from verb to noun (it happens to coincide with the first person present of the verb *contestar* 'to answer').

[6] https://www.corpusdelespanol.org/

[7] These pieces of research use t-score rather than frequency, but the rankings yielded by both of them are strikingly similar.

[8] This threshold was established in order to discard possible happaxes in the reference corpus. The threshold was low in order to have as many data as possible to predict the level of each text.

encompasses quite a variety of lexical combinations, ranging from support verb constructions (e.g. **make/do the homework*), to idiosyncratic combinations where the collocate has a very restricted applicability (*leap year*). For this process our annotator had a list of candidates and could optionally check their context in the corpus by means of a link.

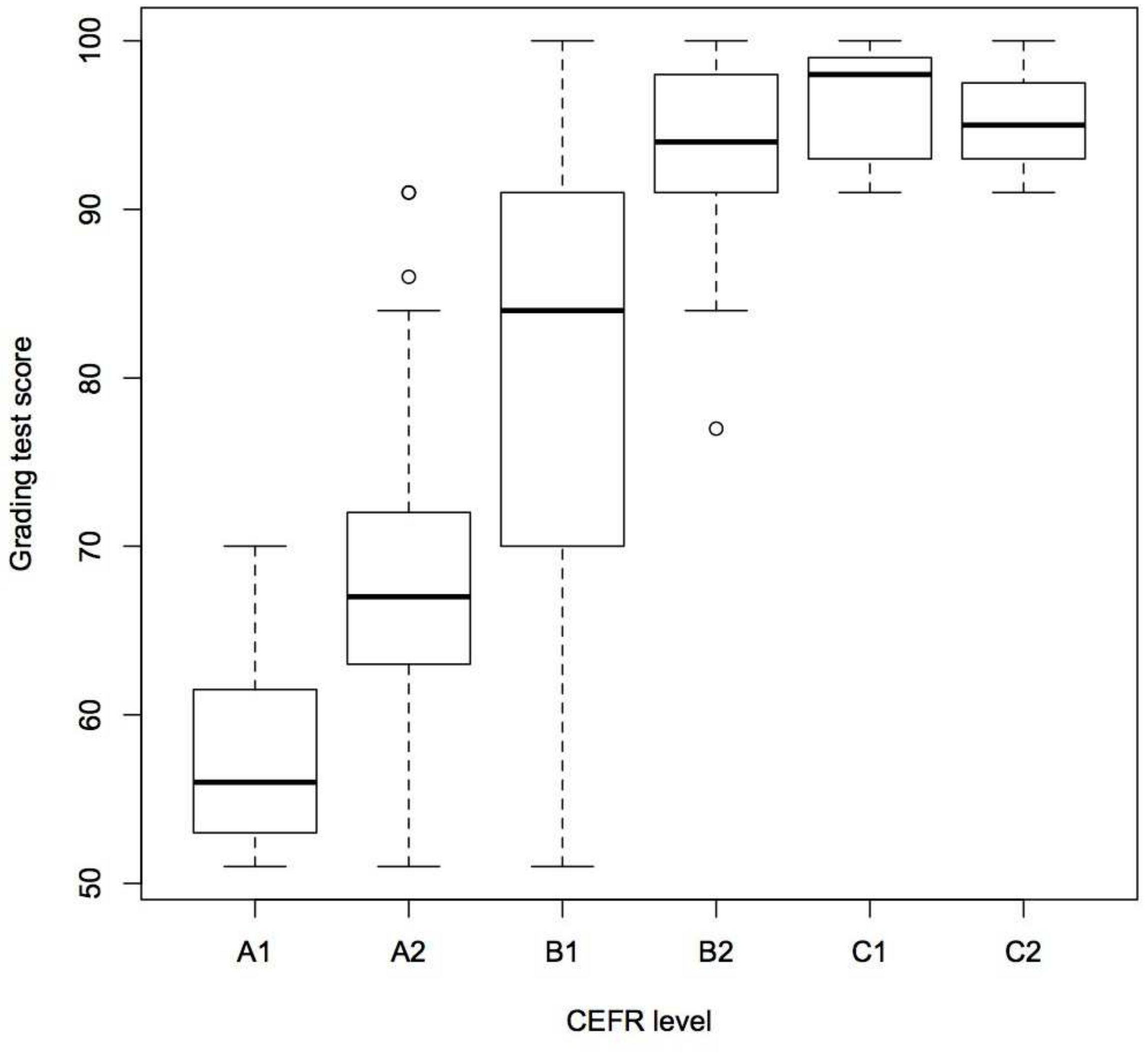

Figure 1: Correspondence between author's scores and CEFR levels of texts. Boxes represent the central 50% of data, the horizontal line in bold is the median and the segments outside boxes are the lowest and highest 25% of data. Outliers are represented by dots.

To evaluate the internal consistency of the annotation we performed an intra-annotator agreement analysis. The collocations of some texts included in the corpus had already been annotated for a previous project, although with a somewhat different procedure. On that occasion, the annotators read the entire texts and annotated their collocations in XML format. The compared samples consisted of 4,867 candidates from the 88 texts that were annotated in both processes. The coincidence was of 87% (Fleiss' kappa = 0.7), a considerable value taking into account the differences between the two annotation processes and the time lapse between them (around five years). This agreement rate also provides indirect evidence on the syntactic parsing quality.

This annotation allowed us to establish a correspondence between the association measures of our sample and the fact that a combination was considered a phraseological collocation by a native speaker of Spanish, thus providing a further selection criterion for candidates inclusion in a dictionary.

4. Results and discussion

All the processes just described resulted in a set of collocation candidates associated to the CEFR level of the texts where they appeared and two association measures taken from their occurrences in the reference corpus.[9] Using these data we examined whether there was any relation between proficiency level and the statistically measured association of candidates appearing in texts graded with such level.

The correspondence between candidate combinations' AMs and CEFR level can be seen in Figures 2 and 3. The first set of data correspond to the whole set of combinations, whereas in Figure 3 the data are mean scores for each text obtained from the association measures of the combinations it contains. As for the data in Figures 2, there seems to be a positive correlation between MI median and CEFR level in all three syntactic patterns, even though a considerable overlap between the different levels is apparent. It is also noticeable that MI values are rather low, particularly in the case of subject-verb combinations, where the medians in all levels fall below 3.

In the case of frequency, the correspondence between level and association scores is much less clear: it could be an inverse correlation in the case of subject-verb combinations, but in the other two syntactic patterns no such tendency emerges and the overlap for verb-object frequency values is almost total.

If we assign an average score to each text based on the AMs of candidates it contains, like in Bestgen and Granger (2014), Granger and Bestgen (2014) and Paquot (2018), a somewhat clearer picture emerges, but the general tendency is similar to that discussed above. In Figure 3 one can observe a clearer tendency for texts of higher levels to obtain higher average values of MI in all three syntactic patterns examined, particularly in adjective-noun combinations. Based on the average scores for these combinations, C1 and C2 are clearly detached from the rest. The MI values for the other two syntactic patterns are generally lower (especially in the case of subject-verb, as before), and some groups divert from the general tendency (namely, B2 in verb-object, which has a lower

[9] In the case of frequency, we used the logarithm of base 10 of the raw frequency—cf .van Heuven et al. (2014).

median than B1, and C1 in subject-verb combinations).

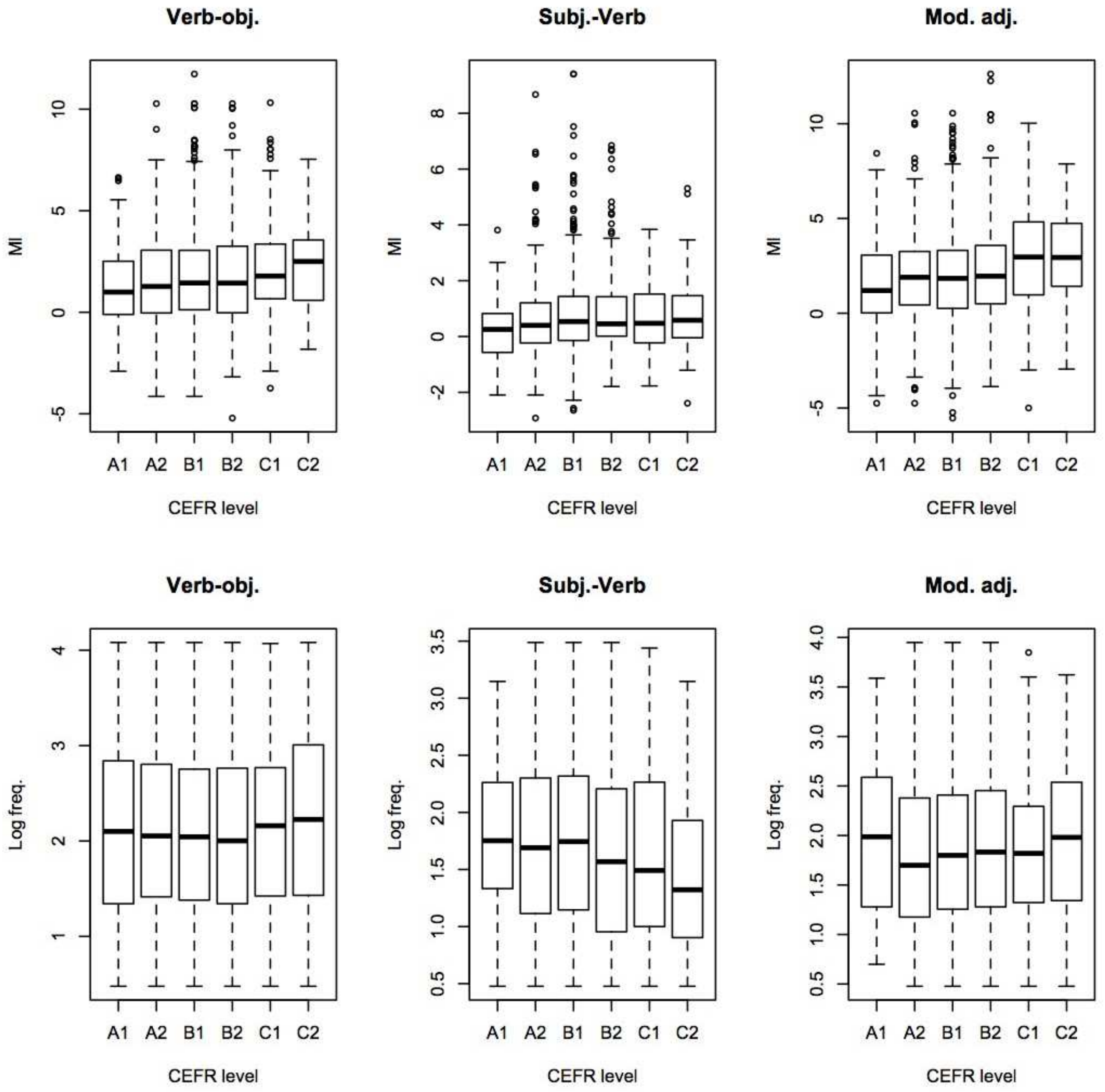

Figure 2: Association measures and CEFR level

With respect to frequency, again no linear progression in text averages can be observed in any of the syntactic patterns studied (although the median of verb-object based averages seems to draw a parabolic line).

In order to establish whether the observed tendencies reached statistical significance, the data were submitted to a generalized linear mixed model analysis.[10] We treated the CEFR levels assigned to the texts as the dependent variable and the mean scores based on AMs as independent variables. Corpus texts were included into the model as random

[10] For this we used R's lme4 package Bates et al. (2015): https://cran.r-project.org/package=lme4.

factors. Using the texts' mean scores based on the three syntactic patterns examined separately did not yield significant results. However, when the mean scores obtained from taking together the AMs of the three syntactic types of combinations were used, significant effects for mean MI and for the interaction between MI and frequency were observed, as can be seen in Table 2. The fixed effects of this model explains 39% of the variance.[11]

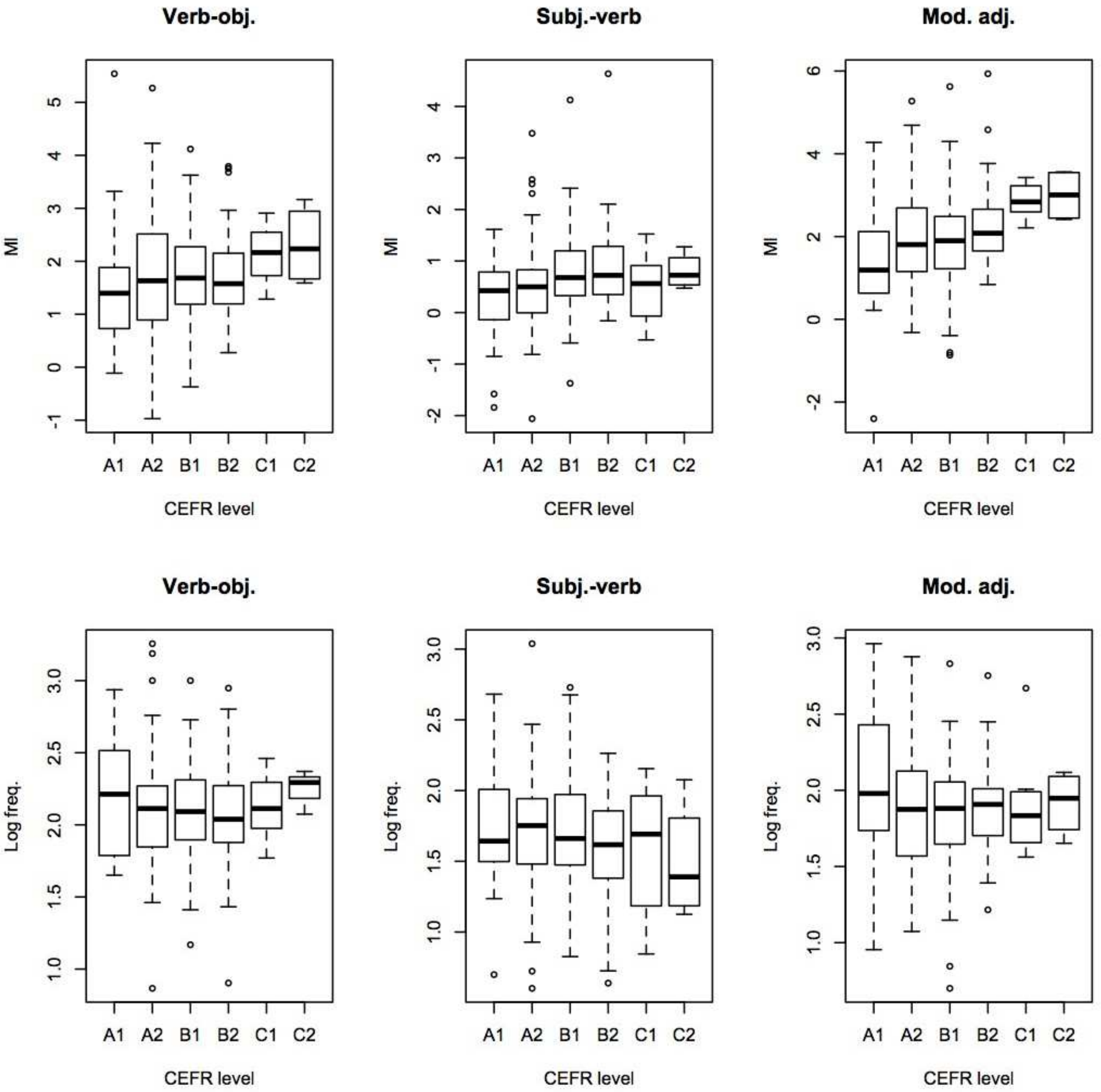

Figure 3: Association measure text means and CEFR level

These results indicate that the raters graded texts containing lexical combinations with higher MI values at more advanced levels. Frequency alone did not have an effect on

[11] R^2 was calculated with R's MuMIn package Barton (2019): https://cran.r-project.org/package=MuMIn

the level assigned to the texts, but it counteracted the effect of MI. This suggests that frequent lexical combinations, even if their members are highly associated (i.e., they have high MI scores), are not perceived as markers of advanced proficiency, at least not so clearly as less frequent combinations with equally high MI scores.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.879	3.937	-0.985	0.32447
mean MI	9.302	3.269	2.845	0.00444 **
mean log freq.	2.386	2.047	1.165	0.24398
mean MI : mean log freq.	-4.028	1.602	-2.514	0.01195 *

Table 2: Fixed effects for the generalised linear mixed model predicting CEFR level. Factors marked with ** are significant at the 0.01 level and those marked with * at the 0.05 level.

All the above indicates that, when it comes to classifying a repertoire of collocations into CEFR proficiency levels, the strength of association between their constituents as measured by MI is more relevant than frequency of co-occurrence. This is in consonance with the findings of Bestgen and Granger (2014) and Paquot (2018) for English. However, given the effect of frequency in the opposite direction, those two dimensions should be somehow combined in such a classification.

The data examined so far refer to lexical combinations that only had to meet two conditions: they must have at least three occurrences in the reference corpus and be instances of one of the three syntactic dependencies referred to above. However, compiling a collocation dictionary that includes all the combinations that met these two requirements is probably not very interesting. Thus, for instance, we have seen the especially low MI values of subject-verb combinations that puts into question the phraseological status of many of the candidates belonging to this pattern. In order to refine the set of candidates, we will now review the data coming from the manual collocation annotation of the learner corpus' sample.

When examining the correlation between the human annotator's criterion and the AMs used here, frequency of co-occurrence shows up as slightly superior to MI in separating good from bad candidates, as can be seen in the precision-recall curves of Figure 4. Even using association measures, our results point to the need of human intervention in compiling a collocation dictionary. Thus, if we wanted to retrieve 80% of phraseological collocations included in the sample by using a log10 frequency value as the cut-off point (which in this case was log10(frequency) ≥ 1.53, or 34 occurrences in raw frequency), the mean precision would be 35%, that is, 65% of candidates would have to be manually discarded.

If we extrapolate these figures to the data extracted from our reference corpus, we will end up with a set of ca. 50,000 candidates, which would eventually yield around 18,000 phraseological collocations.

To gauge what kind of collocation candidates would be extracted for each CEFR level using frequency and MI thresholds, we have used the sextiles corresponding to the MI values of the reference corpus data as cut-off points and extracted the ten best candidates for each level. In the case of A1 through B2 levels the candidates were those with the highest frequencies, whereas for C1 and C2 we extracted those with medium frequencies, in order to reflect somehow the negative interaction between frequency and MI. The results can be seen in Table 3. For the sake of clarity, we occasionally have used inflectional variants different from the lemma form.

A1	tener tiempo 'have time'; tener cosa(s) 'have thing(s)'; tener vida 'have life; tener dinero 'have money'; tener trabajo 'have [a] job'; tener poder 'have power'; (la) gente tiene/tenía/etc. 'people have'; tener punto 'have point(s)?'; tener día 'have day'; tener nombre 'have name';
A2	tener idea 'have idea'; tener relación 'have relationship'; tener posibilidad 'have possibility'; tener opción 'have option'; ver cosa(s) 'see thing(s)'; tener efecto 'have an effect'; tener experiencia 'have experience'; dar vida 'give life'; persona tiene/tenía/etc. 'people have'; tener valor 'have value'; dar tiempo 'give time'
B1	hacer cosa(s) 'do thing'; tener problema(s) 'have problem'; hacer tiempo 'lit. make time, time ago'; tener razón 'be right|have reason'; tener sentido 'have sense|make sense'; tener derecho 'have right'; tener suerte 'have luck'; tener gana(s) 'have desire'; tener oportunidad 'have opportunity'; tener año(s) 'have year'; tener hijo(s) 'have children'
B2	hacer falta 'need, lit. make lack'; mismo tiempo 'same time'; mayor parte 'most of'; gran parte 'large part'; dar paso(s) 'take step(s), lit. give step(s)'; llevar tiempo 'take time, lit. carry time'; ver película 'watch film'; decir cosa(s) 'say things'; gran cantidad 'large quantity'; dar oportunidad 'give opportunity'; hacer daño 'do harm'
C1	desarrollar trama 'develop plot'; transformación profunda 'deep transformation'; volcán alto 'high volcano'; añadir aceite 'add oil'; alojamiento web 'web hosting'; provocar aparición 'cause apparition'; artista extranjero 'foreign artist'; respetar autor 'respect author'; escuchar banda 'listen (to a) band'; ganar batalla 'win (the) battle'
C2	aumento considerable 'considerable increase'; pedir auxilio 'call for help'; bebida gaseosa 'soda'; coger bici 'take (the) bike'; beber café 'drink coffee'; célula cerebral 'brain cell'; certificado digital 'digital certificate'; comida casera 'home-cooked food'; compañía aseguradora 'insurance company'; sintetizar concepto 'sum up (a) concept'

Table 3: Samples of candidates by CEFR level.

Candidates with low MI and high frequency, i.e., those corresponding to A1 and A2 levels, tend to be support verb constructions with one of the most frequent verbs in

Spanish (*tener* 'have'). This is in keeping with what the PCIC proposes. Thus, the most common verbs occurring in multiword expressions at levels A1 and A2 are *hacer* 'do, make' and *tener* 'have' (in addition to *ser* 'be'). The sample here only includes the ten most frequent candidates, and is not very informative about other types of combinations (particularly, noun+adjective), which are less frequent and more scarce. It is at C1 and C2 levels (for which we took samples of medium frequency) where noun+adjective combinations start to appear regularly. Another issue is the presence of some free combinations (*tener cosa(s)* 'have things') seemingly not very interesting for learners, as well as combinations hardly recognisable out of context (*tener punto(s)* 'have points?, have a score?';).

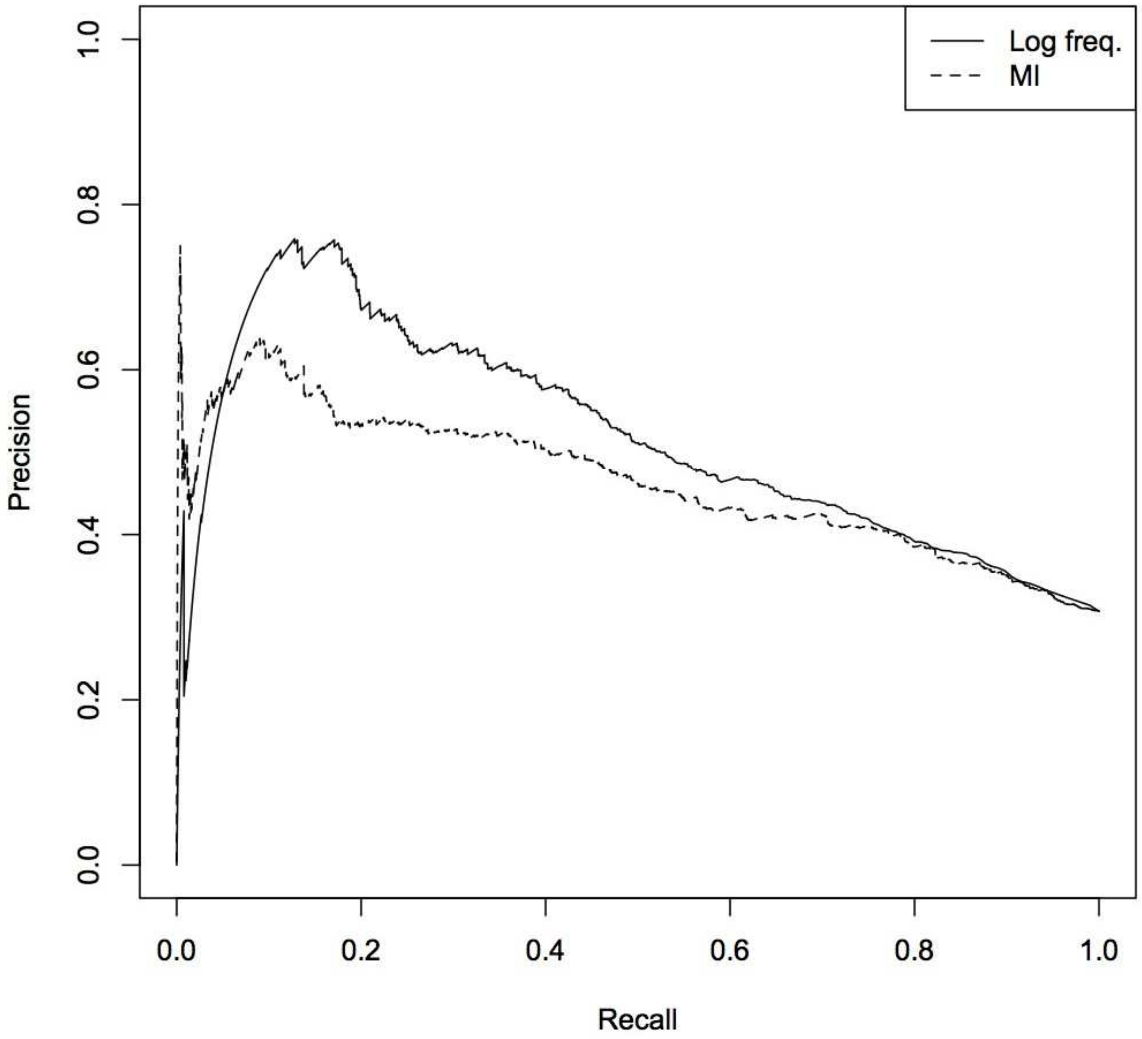

Figure 4: Precision-recall curves for candidates annotated as collocations

As for B1 and B2 levels, although candidate combinations with *tener* are also predominant here, they seem more interesting than those at lower levels. Among them, however, there are some idioms (e.g. *hacer falta* 'be necessary') that require a different lexicographic treatment than collocations. Finally, the candidates in levels C1 and C2 are more variegated: there are more adjective-nouns combinations, although they also include free (e.g. *añadir aceite* 'add oil') and non-compositional combinations (*alojamiento web* 'web-hosting').

These observations call for a manual revision of candidates when compiling a dictionary: uninteresting free combinations probably should be excluded, non-compositional combinations should be distinguished from the rest in order to give them a different lexicographic treatment, etc. Notwithstanding, pre-processing using AMs seems a valuable guiding principle for selection and grading. As far as collocation selection is concerned, using a frequency cut-off point or examining only the n-best candidates ranked by frequency can alleviate the task of lexicographers, since as seen in Fig. 4 different values of AMs are associated with different precision rates. With respect to grading, we have proven that MI has an effect on the CEFR level given by raters.

5. Conclusion

This paper explored the use of association measures to extract collocations with a view to populating a dictionary of Spanish graded by CEFR levels. When it comes to grading collocations, much like in the case of single words, frequency seems in principle a reasonable criterion to determine the sequence of vocabulary presentation in curricula: giving priority to high-frequency lexical elements provides learners with valuable knowledge, both in terms of comprehension and production, given the high coverage rates of these elements.

However, the co-occurrence frequency of lexical combinations is not a good predictor of the proficiency level of learners' texts. In this respect, MI has shown up as clearly superior. This finding is in line with what Granger and Bestgen (2014) and Paquot (2018) observe regarding the text quality of English learners. In consequence, future lexicographic ventures should take into account MI when it comes to grading lexical combinations.

Frequency, in turn, seems to perform slightly better than MI in distinguishing collocations from other types of lexical combinations (free, non-compositional) as identified by human annotators following phraseological criteria. This is at odds with some previous research (Ellis et al., 2008) and probably deserves further investigation. A possible reason is that here we used candidates in a syntactic relationship, rather than candidates within a given text span, in contrast to Ellis et al. (2008) (cf. Garcia et al., 2019, for similar results with a native sample).

Whereas the two association measures examined can ease the task of lexicographers by promoting collocational candidates (frequency) and providing a sequencing criterion (MI), they cannot guarantee a completely automated process with high quality results.

This study presented an initial approach that opens up further lines of research, starting with replications with more balanced data in terms of the representation of the different CEFR levels in the corpus—not an easy task given the difficulty to come by Spanish learner corpora of sizes comparable to those pertaining to other genres. Additionally, we have dealt with only two AMs, due to their spread use in related studies. However, a plethora of lexical AMs has been proposed (Pecina, 2010). It will be interesting, therefore, to study the correspondence between those measures and learners' proficiency in future studies.

6. Acknowledgements

Marcos García Salido is the recipient of a post-doctoral grant from Xunta de Galicia (ED481D 2017/009), and Marcos Garcia's research is funded by a Juan de la Cierva incorporación grant (IJCI-2016-29598) and a 2017 Leonardo Grant for Researchers and Cultural Creators (BBVA Foundation). This research was also supported by a project funded by Ministerio de Economía, Industria y Competitividad (FFI2016-78299-P).

7. References

Alonso Ramos, M., Roberto Carlini, J. C. F., Orol González, A., Vincze, O. & Wanner, L. (2015). Towards a learner need-oriented second language collocation writing assistant. In F. Helm, L. Bradley, M. Guarda & S. Thouësny (eds.) *Critical CALL–Proceedings of the 2015 EUROCALL Conference*, 2015. p. 16. https://reference.research-publishing.net/publication/chapters/978-1-908416-29-2/304.pdf.

Alonso-Ramos, M. (2004). Diccionario de colocaciones del español. http://www.dicesp.com/.

Alvar Ezquerra, M. (2005). La frecuencia léxica y su utilidad en la enseñanza del español como lengua extranjera. In M. A. C. Carballo, O. C. Moya, J. M. G. Platero & J. P. M. Gutiérrez (eds.) *Actas del XV Congreso Internacional de ASELE.* pp. 19–39. http://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/15/15_0017.pdf.

Barton, K. (2019). Package 'MuMIn'. R Package Version 1.43.6. https://cran.r-project.org/package=MuMIn.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), pp. 1–48.

Benson, M., Benson, E. & Ilson, R. (1986). *The BBI combinatory dictionary of English: A guide to word combinations.* John Benjamins Publishing.

Bestgen, Y. & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26, pp. 28–41. http://dx.doi.org/10.1016/j.jslw.2014.09.004.

Bosque, I. (2004). *REDES. Diccionario combinatorio del español contemporáneo.* Madrid: SM.

Bosque, I. (2006). *Diccionario combinatorio práctico del español contemporáneo.*

Madrid: SM.

Capel, A. (2010). A1 - B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1(01), p. e3. http://dx.doi.org/10.1017/S2041536210000048.

Capel, A. (2012). Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, 3(June 2012), p. e1. http://dx.doi.org/10.1017/S2041536212000013.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press.

Ellis, N. C., Simpson-Vlach, R. & Maynard, C. (2008). Formulaic language in native and second-language speakers. *TESOL Quarterly*, 42(3), pp. 375–396.

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

Garcia, M. & Gamallo, P. (2015). Yet Another Suite of Multilingual NLP Tools. In J.P.L. J. L. Sierra-Rodríguez & A. Simões (eds.) *Languages, Applications and Technologies. Communications in Computer and Information Science.* Cham: Springer, pp. 65–75.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019). A comparison of statistical association measures for identifying dependency-based collocations in various languages. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWEWN 2019).* Florence: Association for Computational Linguistics.

García Salido, M. & Alonso Ramos, M. (2017). Asignación de niveles de aprendizaje a las colocaciones del Diccionario de Colocaciones del español. *Revista Signos*, 51(97), pp. 153–174.

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

Granger, S. & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *IRAL - International Review of Applied Linguistics in Language Teaching*, 52(3), pp. 229–252.

Higueras García, M. (2017). Pedagogical principles for the teaching of collocations in the foreign language classroom. In S. Torner & E. Bernal (eds.) *Collocations and other lexical combinations in Spanish: theoretical, lexicographical and applied perspectives.* London & New York: Routledge, pp. 250–266.

Lozano, C. & Mendikoetxea, A. (2013). Learner corpora and second language acquisition. The design and collection of CEDEL2. In A. Díaz-Negrillo, P. Thompson & N. Ballier (eds.) *Automatic Treatment and Analysis of Learner Corpus Data.* Amsterdam/Philadelphia: John Benjamins, pp. 65–100.

Martinez, R. (2013). A framework for the inclusion of multi-word expressions in ELT. *ELT Journal*, 67(April), pp. 184–198.

Mel'čuk, I. (2012). Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology*, 3, pp. 31–56.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nesselhauf, N. (2004). *Collocations in a Learner Corpus*. Amsterdam/Philadelphia: John Benjamins.

Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In N. Calzolari, K. Choukri, T. Declerck, M.U. Dogan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis (eds.) *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2012)*. European Language Resources Association (ELRA), pp. 2473–2479. http://dblp.uni-trier.de/db/conf/lrec/lrec2012.html#PadroS12.

Paquot, M. (2018). Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights From a Study of EFL Learners' Use of Statistical Collocations. *Language Assessment Quarterly*, 15(1), pp. 29–43.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2), pp. 137–158.

Spina, S. (2016). Learner corpus research and phraseology in Italian as a second language: The case of the DICI-A, a learner dictionary of Italian collocations. In B. Sanromán Vilas (ed.) *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching*. Helsinki: Société Néophilologique de Helsinki, pp. 219–244.

Straka, M., Hajic, J. & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), pp. 1659–1666.

van Heuven, W. J., Mandera, P., Keuleers, E. & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67(6), pp. 1176–1190.

Vincze, O., García-Salido, M., Orol, A. & Alonso-Ramos, M. (2016). A corpus study of Spanish as a Foreign Language learners' collocation production. In M. Alonso-Ramos (ed.) *Spanish Learner Corpus Research*. Amsterdam/Philadelphia: John Benjamins, pp. 299–331. https://benjamins.com/catalog/scl.78.11vin.