Classification of Lexical Collocation Errors in the Writings of Learners of Spanish

Sara Rodríguez-Fernández
Universitat Pompeu Fabra

Roberto Carlini
Universitat Pompeu Fabra

Leo Wanner
ICREA and
Universitat Pompeu Fabra

{sara.rodriguez.fernandez, roberto.carlini, leo.wanner}@upf.edu

Abstract

It is generally acknowledged that *collocations* in the sense of idiosyncratic word co-occurrences are a challenge in the context of second language learning. Advanced miscollocation correction is thus highly desirable. However, state-of-the-art "collocation checkers" are merely able to detect a possible miscollocation and then offer as correction suggestion a list of collocations of the given keyword retrieved automatically from a corpus. No more targeted correction is possible since state-of-the-art collocation checkers are not able to identify the type of the miscollocation. We suggest a classification of the main types of lexical miscollocations by US American learners of Spanish and demonstrate its performance.

1 Introduction

In the second language learning literature, it is generally acknowledged that it is in particular idiosyncratic word co-occurrences of the kind *take* [*a*] *walk*, *make* [*a*] *proposal*, *pass* [*an*] *exam*, *weak performance*, *hard blow*, etc. that make language learning a challenge (Granger, 1998; Lewis, 2000; Nesselhauf, 2004; Nesselhauf, 2005; Lesniewska, 2006; Alonso Ramos et al., 2010). Such co-occurrences (in lexicography known as "collocations") are language-specific. For instance, in Spanish, you 'give a walk' (*dar* [*un*] *paseo*), while in French and German you 'make' it (*faire* [*une*] *promenade* / [*einen*] *Spaziergang machen*). In English you *take a step*, while in German you 'make' it ([*einen*] *Schritt machen*) and in Spanish you 'give' it (*dar* [*un*] *paso*). In English, you can *hold* or *give* [*a*] *lecture*, in Spanish you 'give' (*dar* [*una*] *clase*), but you do not 'hold' it, and in German you 'hold' it ([*eine*] *Vorlesung halten*), but do not 'give' it. And so on.

Several proposals have been put forward for how to verify automatically whether a collocation as used by a language learner is correct or not and, in the case that it is not, display a list of potential collocations of the keyword (*walk*, *step*, and *lecture* above) of the assumingly incorrect collocation. For instance, a Spanish learner of English may use **approve* [*an*] *exam* instead of *pass* [*an*] *exam*. When this miscollocation is entered, e.g., into the MUST collocation checker[1] for verification, the program suggests (in this order) *pass exam*, *sit exam*, *take exam*, *fail exam*, and *do exam* as possible corrections. That is, the checker offers all possible <verb> + *exam* collocations found in a reference corpus or dictionary. However, the display of a mere list of correct collocations of a given keyword is unsatisfactory for learners since they are left alone with the problem of picking the right one among several (potentially rather similar) choices. On the other hand, no further restriction of the list of correction candidates or any meaningful reordering is possible because the collocation checker has no knowledge about the type of the error of the miscollocation.

In order to improve the state of affairs, and be able to propose a more targeted correction, we must be able to identify the type of error of the collocation proposed by the learner (and thus also the meaning the learner intended to express by the miscollocation). While this seems hardly feasible with isolated collocations submitted by a learner for verification (as above), error type recognition in the writings of learners is more promising. Such an error type recognition procedure is taken for granted in grammar checkers, but is still absolutely unexplored in collocation checkers. In what follows, we outline how some of the most prominent errors in collocations identified in the writings of US American students learning Spanish can be

[1] http://miscollocation-richtrf.rhcloud.com/

Proceedings of Recent Advances in Natural Language Processing, pages 529–536,
Hissar, Bulgaria, Sep 7–9 2015.

classified with respect to a given collocation error typology.

2 Background on Collocations and Collocation Errors

Given that the notion of collocation has been discussed and interpreted in lexicology from different angles, we first clarify our usage of the term. Then, we outline the miscollocation typology that underlies our classification.

2.1 On the Nature of Collocations

The term "collocation" as introduced by Firth (1957) and cast into a definition by Halliday (1961) encompasses the statistical distribution of lexical items in context: lexical items that form high probability associations are considered collocations. It is this interpretation that underlies most works on automatic identification of collocations in corpora; see, e.g., (Choueka, 1988; Church and Hanks, 1989; Pecina, 2008; Evert, 2007; Bouma, 2010). However, in contemporary lexicography and lexicology, an interpretation that stresses the idiosyncratic nature of collocations prevails. According to Hausmann (1984), Cowie (1994), Mel'čuk (1995) and others, a collocation is a binary idiosyncratic co-occurrence of lexical items between which a direct syntactic dependency holds and where the occurrence of one of the items (the *base*) is subject of the free choice of the speaker, while the occurrence of the other item (the *collocate*) is restricted by the base. Thus, in the case of *take* [*a*] *walk*, *walk* is the base and *take* the collocate, in the case of *high speed*, *speed* is the base and *high* the collocate, etc. It is this understanding of the term "collocation" that we find reflected in general public collocation dictionaries and that we follow in our work since it seems most useful in the context of second language acquisition. However, this is not to say that the two main interpretations of the term "collocation", the distributional and the idiosyncratic one, are disjoint, i.e., necessarily lead to a different judgement with respect to the collocation status of a word combination. On the contrary: two lexical items that form an idiosyncratic co-occurrence are likely to occur together in a corpus with a high value of *Pointwise Mutual Information* (PMI) (Church and Hanks, 1989):

$$PMI = \log\left(\frac{P(a\cap b)}{P(a)P(b)}\right) = \log\left(\frac{P(a|b)}{P(a)}\right) = \log\left(\frac{P(b|a)}{P(b)}\right) \quad (1)$$

The PMI indicates that if two variables *a* and *b* are independent, the probability of their intersection is the product of their probabilities. A PMI equal to 0 means that the variables are independent; a positive PMI implies a correlation beyond independence; and a negative PMI signals that the co-occurrence of the variables is lower than the average. Two lexemes are thus considered to form a collocation when they have a positive PMI, i.e., they are found together more often that this would happen if they would be independent variables.

PMI has been a standard collocation measure throughout the literature since Church and Hank's proposal in 1989. However, a mere use of PMI or any similar measure neglects that the lexical dependencies between the base and the collocate are not symmetric (recall that PMI is commutative, i.e., $PMI(a,b) = PMI(b,a)$). Only a few studies take into consideration the asymmetry of collocations; see, e.g., Gries (2013), who proposes an asymmetric association measure, ΔP, and Carlini et al. (2014), who propose an assymmetric normalization of PMI; see Eq. (2). In our work, we use Carlini et al. (2014)'s asymmetric $NPMI_C$.

$$NPMI_C = \frac{PMI(collocate,base)}{-log(p(collocate))} \quad (2)$$

2.2 Typology of Collocation Errors

Alonso Ramos et al. (2010) proposed a detailed three-dimensional typology of collocation errors. The first dimension defines which element of the collocation (the base or the collocate) is erroneous or whether it is the collocation as a whole. The second (descriptive) dimension details the type of error that was produced. Three different global types are distinguished: register, lexical, and grammatical. The third dimension, finally, details the possible interpretation of the origin of the error (e.g., calque from the native language of the learner, analogy to another common collocation, etc.). In the experiments presented in this paper, we focus on the lexical branch of the descriptive dimension.

Lexical errors are divided into five different types; the first two affect either the base or the collocate; the other three the collocation as a whole:[2]

[2]Given that we work on a Spanish learner corpus, the examples of miscollocations are in Spanish. The consensual-

1. *Substitution errors*: Errors resulting from an inappropriate choice of a lexical unit that exists in the language as either base or collocate. This is the case, e.g., with **realizar una meta* 'to reach a goal', lit. 'to make, to carry out a goal', where both the base and the collocate are existing lexical units in Spanish, but the correct collocate *alcanzar*, lit. 'to achieve' has been substituted by *realizar*.

2. *Creation errors*: Errors resulting from the use of a non-existing (i.e., "created" or invented) lexical unit as the base or as the collocate. An example of this type of error is **estallar confrontamientos*, instead of *estallar confrontaciones*, lit. '(make) explode a confrontation', where the learner has used the non-existing form *confrontamientos*.

3. *Synthesis errors*: Errors resulting from the use of a non-existing lexical unit instead of a collocation, as, for instance, **escaparatear*, instead of *ir de escaparates* 'to go window-shopping'.

4. *Analysis errors*: Errors that are inverse to synthesis errors, i.e., that result from the use of an invented collocation instead of a single lexical unit expression. An example of this type of error is **sitio de acampar* 'camping site', which in Spanish would be better expressed by the lexical unit *camping*.

5. *Different sense errors*: Errors resulting from the use of a correct collocation, but with meaning different from the intended one. An example of this type of error is **el próximo día*, instead of *el día siguiente* 'the next day'.

Our studies show that 'Substitution', 'Creation' and 'Different sense' errors are the most common types of miscollocations. In contrast, learners tend to make rather few 'Synthesis' and 'Analysis' errors. Therefore, given that 'Synthesis' errors are not comparable to any other error class, we decided not to consider them at this stage of our work. 'Analysis' errors show in their appearance a high similarity to 'Substitution' errors, such that they could be merged with them without any major distortion of the typology. Therefore, we deal below with miscollocation classification with respect to three lexical error classes: 1. 'Extended Substitution', 2. 'Creation', and 3. 'Different Sense'.

ized judgement whether a given collocation is a miscollocation or a correct collocation has in all cases been made by a team of lexicographers who are native speakers of Peninsular Spanish.

3 Towards Automatic Collocation Error Classification

In corpus-based linguistic phenomenon classification, it is common to choose a supervised machine learning method that is then used to assign any identified phenomenon to one of the available classes. In the light of the diversity of the linguistic nature of the collocation errors and the widely diverging frequency of the different error types, this procedure seems not optimal for miscollocation classification. A round of preliminary experiments confirmed this assessment. It is more promising to target the identification of each collocation error type separately, using for each of them the identification method that suits its characteristics best. Furthermore and as a matter of fact, it cannot be excluded that a miscollocation may contain more than one type of error. Thus, it may contain an error in the base and another error in the collocate, or it might have a lexical and a grammatical error or two lexical errors (one per element) at the same time. An example of a collocation containing two lexical errors is *afecto malo* 'bad effect', where both the base and the collocate are incorrect. *Afecto* 'affect' is chosen instead of *efecto* 'effect', and *malo* 'bad' instead of *nocivo* 'damaging'.

In what follows, we describe the methods that we use to identify miscollocations of the three types that we target. All of these methods perform a binary classification of all identified incorrect collocations as 'of type X' / 'not of type X'. The methods for the identification of 'Extended substitution' and 'Creation' errors receive as input the incorrect collocations (i.e., grammatical, lexical or register-oriented miscollocations) recognized in the writing of a language learner by a collocation error recognition program[3], together with their sentential contexts. The method for the recognition of 'Different sense' errors receives as input 'different sense' errors along with the correct

[3]Since in our experiments we focus on miscollocation classification, we use as "writings of language learners" a learner corpus in which both correct and incorrect collocations have been annotated manually and revised by different annotators. Only those instances for which complete agreement was found were used for the experiments.

collocations identified in the writing of the learner.

Extended Substitution Error Classification. For the classification of incorrect collocations as 'extended substitution error' / 'not an extended substitution error', we use supervised machine learning. This is because 'extended substitution' is, on the one side, the most common type of error (such that sufficient training material is available), and, on the other side, very variant (such that it is difficult to be captured by a rule-based procedure). After testing various ML-approaches, we have chosen the Support Vector Machine (SMO) implementation from the Weka toolkit (Hall et al., 2009).[4]

Two different types of features have been used: lexical features and co-occurrence (or PMI-based) features. The lexical features consist of the lemma of the collocate and the bigram made up of the lemmas of the base and collocate. The PMI-based features consist of: $NPMI_C$ of the base and the collocate, $NPMI_C$ of the hypernym of the base and the collocate, $NPMI$ of the base and its context, and $NPMI$ of the collocate and its context, considering as context the two immediate words to the left and to the right of each element. Hypernyms were taken from the Spanish WordNet; $NPMI$s and $NPMI_C$s were calculated on a 7 million sentences reference corpus of Spanish.

Creation Error Classification. For the detection of creation errors among all miscollocations, we have designed a rule-based algorithm that uses linguistic (lexical and morphological) information; see Algorithm 1.

If both elements of a collocation under examination are found in the reference corpus (RC) with a sufficient frequency ($\geq$50 for our experiments), they are considered valid tokens of Spanish, and therefore 'Not creation' errors. If one of the elements has a low frequency in the RC ($<$50), the algorithm continues to examine the miscollocation. First, it checks whether a learner used an English word in a Spanish sentence, considering it as a 'transfer Creation error'. If this is not the case, it checks whether the gender suffix is wrong, considering it as a 'gender Creation error', as in, e.g., **hacer regala* instead of *hacer regalo*, lit. 'make present'. This is done by alternating the gender suffix and checking the resulting token in the RC.

[4]Weka is University of Waikato's public machine learning platform that offers a great variety of different classification algorithms for data mining.

Algorithm 1: Creation Error Classification

Given a collocation '$b + c$' that is to be verified
do

if b_L,$c_L \in$ RC
// with 'b_L'/'c_L' as lemmatized base/collocate
  and freq('b_L') > 50
  and freq('c_L') > 50
  then echo "Not a creation error"
else if $b_L \vee c_L \in$ English dictionary
  then echo "Creation error (Transfer)"
else if check_gender(b_L) = false
  then echo "Creation error (Incorrect gender)"
else if check_affix(b_r) || check_affix(c_r)
// with 'b_r'/'c_r' as stems of base/collocate
  then echo: "Creation error (Incorrect derivation)"
else if check_ortography(b_L) || check_ortography(c_L)
  then echo "Not a creation error (Ortographic)"
else if freq('b_L') > 0 **or** freq('c_L') > 0
  then echo "Not a creation error"
else
echo "Creation error (Unidentified)"

If no gender-influenced error could be detected, the algorithm checks whether the error is due to an incorrect morphological derivation of either the base or the collocate — which would imply a 'derivation Creation error', as in, e.g. **ataque terroríst<u>ico</u>* instead of *ataque terror<u>ista</u>* 'terrorist attack'. For this purpose, the stems of the collocation elements are obtained and expanded by the common nominal / verbal derivation affixes of Spanish to see whether any derivation leads to the form used by the learner. Should this not be the case, the final check is to see whether any of the elements is misspelled and therefore we face a 'Not creation error'. This is done by calculating the edit distance from the given forms to valid tokens in the RC.

In the case of an unsuccessful orthography check, we assume a 'Creation' error if the frequency of one of the elements of the miscollocation is '0', and a 'Not creation' error for element frequencies between '0' and '50'.

Different Sense Error Classification. Given that 'Different Sense Errors' capture the use of correct collocations in an inappropriate context, the main strategy for their detection is to compare the context of a learner collocation with its prototypical context. The prototypical context is represented by a centroid vector calculated using the lexical contexts of the correct uses of the collocation found in the RC.

The vector representing the original context is compared to the centroid vector in terms of cosine

similarity; cf. Eq. (3).

$$sim(A, B) = \frac{A \cdot B}{\|A\| \, \|B\|} \quad (3)$$

A specific similarity threshold must be determined in order to discriminate correct and incorrect uses. In the experiments we carried out so far, 0.02543 was empirically determined as the best fitting threshold. However, further research is needed to design a more generic threshold determination procedure.

4 Experiments

In this section, we first describe the experiment set up and present then the results of the experiments.

4.1 Experiment Setup

For our experiments, we use a fragment of the Spanish Learner Corpus CEDEL2 (Lozano, 2009), which is composed of writings of learners of Spanish whose first language is American English. The writings have an average length of 500 words and cover different genres. Opinion essays, descriptive texts, accounts of some past experience, and letters are the most common of them. The levels of the students range from 'low-intermediate' to 'advanced'. In the fragment of CEDEL2 (in total, 517 writings) that we use (our working corpus), both the correct and incorrect collocation occurrences are tagged.[5] As stated above, collocations were annotated and revised, and only those for which a general agreement regarding their status was found, were used for the experiments.

Table 1 shows the frequency of the correct collocations and of the five types of lexical miscollocations in our working corpus. The numbers confirm our decision to discard synthesis miscollocations (there are only 9 of them – compared to, e.g., 565 substitution miscollocations) and to merge analysis miscollocations (19 in our corpus) with substitution miscollocations.[6]

To be able to take the syntactic structure of collocations into account, we processed CEDEL2 with Bohnet (2010)'s syntactic dependency parser[7].

As a reference corpus, we used a seven million sentence corpus, from Peninsular Spanish newspaper material. The reference corpus was also processed with Bohnet (2010)'s syntactic dependency parser.

Class	# Instances
Correct collocations	3245
Analysis errors	19
Substitution errors	565
Creation errors	69
Synthesis errors	9
Different sense errors	48

Table 1: Number of instances of the different types of lexical errors and correct collocations in our working corpus.

4.2 Results of the Experiments

Table 2 shows the performance of the individual collocation error classification methods. In the '+' column of each error type, the accuracy is displayed with which our algorithms correctly detect that a miscollocation belongs to the error type in question; in the '−' column, the accuracy is displayed with which our algorithms correctly detect that a miscollocation does not belong to the corresponding error type.

	'Ext. subst'		'Creation'		'Diff. sense'	
	+	-	+	-	+	-
Baseline	0.395	0.902	0.391	0.986	0.5	0.453
Our model	0.832	0.719	0.681	0.942	0.583	0.587

Table 2: Error detection performance. The lower row displays the achieved accuracy.

To assess the performance of our classification, we use three baselines, one for each type of error. To the best of our knowledge, no other state-of-the-art figures are available with which we could compare its quality further. For the 'Extended substitution' miscollocation classification, we use as baseline a simplified version of the model, trained only with one of our lexical features, namely bigrams made up of the lemmas of the base and

[5]The tagging procedure has been carried out manually by several linguists. The first phase of it was already carried out by (Alonso Ramos et al., 2011). We carried on the tagging work by Alonso Ramos et al. to have for our experiments a corpus of a sufficient size.

[6]Recall that we argued that synthesis miscollocations are too different from the other types of errors to be merged with any other type.

[7]Processing tools' performance on non-native texts is lower than on texts written by natives. We evaluated the performance of the parser on our learner corpus and obtained the following results: LAS:88.50%, UAS:87.67%, LA:84.54%.

the collocate of the collocation. For 'Creation' miscollocation classification, the baseline is an algorithm that judges a miscollocation to be of the type 'Creation' if either one of the elements (the lemma of the base or of the collocate) or both elements of the miscollocation are not found in the reference corpus. Finally, for the 'Different sense' miscollocation classification, we take as baseline an algorithm that, given a bag of the lexical items that constitute the contexts of the correct uses of a collocation in the RC, judges a collocation to be a miscollocation of the 'Different sense' type, if less than half of the lexical items of the context of this collocation in the writing of the learner is not found in the reference bag.

5 Discussion

Before we discuss the outcome of the experiments, let us briefly make some generic remarks on the phenomenon of a collocation in the experiments.

5.1 The Phenomenon of a Collocation

The decision whether a collocation is correct or incorrect is not always straightforward, even for native expert annotators. Firstly, a certain number of collocations was affected by spelling and inflective errors. Consider, e.g., *tomamos cervesas* 'we drank beer', instead of *cervezas*; *sacque una mala nota* 'I got a bad mark', where *saqué* is the right form, or *el dolor disminúe* 'the pain decreases', instead of *disminuye*. In such cases, we assume that these are orthographical or morphological mistakes, rather than collocational ones. Therefore, we consider them to be correct. On the other hand, collocations may also differ in their degree of acceptability. Consider, e.g., *asistir a la escuela, tomar una fotografía* o *mirar la televisión*. Collocations that were doubtful to one or several annotators were looked up in th RC. If their frequency was higher than a certain threshold, they were annotated as correct. Otherwise, they were considered incorrect. From the above examples, *asistir a la escuela* was the only collocation considered as correct after the consultation of the RC.

5.2 The Outcome of the Experiments

The performance figures show that the correct identification of 'Different sense' miscollocations is still a challenge. With an accuracy somewhat below 60% for both the recognition of 'Different sense' miscollocations and recognition of 'Correctly used' collocations, there is room for improvement. Our cosine-measure quite often leads to the classification of correct collocations as 'Different sense' miscollocations (cf., e.g., *ir en coche* 'go by car' , *tener una relación* 'have a relationship', *tener impacto* 'have impact', *tener capacidad* 'have capacity') or classifies 'Different sense' errors as correctly used collocations, such as *gastar el tiempo* (intended *pasar el tiempo* 'spend time' or *tener opciones* instead of *ofrecer posibilidades* 'offer possibilities'. This shows the limitations of an exclusive use of lexical contexts for the judgement whether a collocation is appropriately used: on the one hand, lexical contexts can, in fact, be rather variant (such that the learner may use a collocation correctly in a novel context), and, on the other hand, lexical contexts do not capture the *situational* contexts, which determine even to a major extent the appropriateness of the use of a given expression. Unfortunately, to capture situational contexts remains a big challenge.

6 Conclusions and Future Work

We discussed a classification of collocation errors made by American English learners of Spanish with respect to the lexical branch of the miscollocation typology presented in Alonso Ramos et al. (2010). The results are very good for two of the three error types we considered, 'Substitution' and 'Creation'. The third type of miscollocation, 'Different sense', is recognized to a certain extent, but further research is needed to be able to recognize it as well as the other two error types. But already with the provided classification at hand, learners can be offered much more targeted correction aids than this is the case with the state-of-the-art collocation checkers. We are now about to implement such aids, which will also offer the classification and targeted correction of grammatical collocation errors (Rodríguez-Fernández et al., 2015), into the collocation learning workbench HARenES (Wanner et al., 2013; Alonso Ramos et al., 2015).

7 Acknowledgements

This work is funded by the Spanish Ministry of Economy and Competitiveness (MINECO), through a predoctoral grant (BES-2012-057036), in the framework of the project HARenES (FFI2011-30219-C02-02).

References

Margarita Alonso Ramos, Leo Wanner, Orsolya Vincze, Gerard Casamayor, Nancy Vázquez, Estela Mosqueira, and Sabela Prieto. 2010. Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 3209–3214, La Valetta, Malta.

Margarita Alonso Ramos, Leo Wanner, Orsolya Vincze, Rogelio Nazar, Gabriela Ferraro, Estela Mosqueira, and Sabela Prieto. 2011. Annotation of collocations in a learner corpus for building a learning environment. In *Proceedings of the Learner Corpus Research 2011 Conference*, Louvain-la-Neuve, Belgium.

Margarita Alonso Ramos, Roberto Carlini, Joan Codina-Filba, Ana Orol, Orsolya Vincze, and Leo Wanner. 2015. Towards a Learner Need-Oriented Second Language Collocation Writing Assistant. In *Proceedings of the EURCALL Conference*, Padova, Italy.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 89–97. Association for Computational Linguistics.

Gossa Bouma. 2010. Collocation extraction beyond the independence assumption. In *Proceedings of the ACL 2010, Short paper track*, Uppsala.

Roberto Carlini, Joan Codina-Filba, and Leo Wanner. 2014. Improving Collocation Correction by ranking suggestions using linguistic knowledge. In *Proceedings of the 3rd Workshop on NLP for Computer-Assisted Language Learning*, Uppsala, Sweden.

Yakov Choueka. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO*, pages 34–38.

Keith Church and Patrick Hanks. 1989. Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the ACL*, pages 76–83.

Anthony Cowie. 1994. Phraseology. In R.E. Asher and J.M.Y. Simpson, editors, *The Encyclopedia of Language and Linguistics, Vol. 6*, pages 3168–3171. Pergamon, Oxford.

Stefan Evert. 2007. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.

John Firth. 1957. Modes of meaning. In J.R. Firth, editor, *Papers in Linguistics, 1934-1951*, pages 190–215. Oxford University Press, Oxford.

Sylviane Granger. 1998. Prefabricated patterns in advanced EFL writing: Collocations and Formulae. In A. Cowie, editor, *Phraseology: Theory, Analysis and Applications*, pages 145–160. Oxford University Press, Oxford.

Stefan Th Gries. 2013. 50-something years of work on collocations: what is or should be next. *International Journal of Corpus Linguistics*, 18(1):137–166.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).

Michael Halliday. 1961. Categories of the theory of grammar. *Word*, 17:241–292.

Franz-Joseph Hausmann. 1984. Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortwendungen. *Praxis des neusprachlichen Unterrichts*, 31(1):395–406.

Justyna Lesniewska. 2006. Collocations and second language use. *Studia Lingüística Universitatis lagellonicae Cracoviensis*, 123:95–105.

Michael Lewis. 2000. *Teaching Collocation. Further Developments in the Lexical Approach*. LTP, London.

Cristóbal Lozano. 2009. CEDEL2: Corpus escrito del español L2. In C.M. Bretones Callejas, editor, *Applied Linguistics Now: Understanding Language and Mind*, pages 197–212. Universidad de Almería, Almería.

Igor Mel'čuk. 1995. Phrasemes in Language and Phraseology in Linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk, and R. Schreuder, editors, *Idioms: Structural and Psychological Perspectives*, pages 167–232. Lawrence Erlbaum Associates, Hillsdale.

Nadja Nesselhauf. 2004. How learner corpus analysis can contribute to language teaching: A study of support verb constructions. In G. Aston, S. Bernardini, and D. Stewart, editors, *Corpora and language learners*, pages 109–124. Benjamins Academic Publishers, Amsterdam.

Nadja Nesselhauf. 2005. *Collocations in a Learner Corpus*. Benjamins Academic Publishers, Amsterdam.

Pavel Pecina. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 54–57, Marrakech.

Sara Rodríguez-Fernández, Roberto Carlini, and Leo Wanner. 2015. Classification of Grammatical Collocation Errors in the Writings of Learners of Spanish. In *Proceedings of the Annual Spanish Computational Linguistics Conference (SEPLN)*, Alicante, Spain.

Leo Wanner, Serge Verlinde, and Margarita Alonso Ramos. 2013. Writing Assistants and Automated Lexical Error Correction: Word Combinatorics. In *Proceedings of eLex 2013: Electronic Lexicography in the 21st Century*, Tallinn, Estonia.