Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 49-56 recibido 27-03-2015 revisado 27-04-2015 aceptado 22-05-2015

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

Clasificación de errores gramaticales colocacionales en textos de estudiantes de español

Sara Rodríguez-Fernández
DTIC, UPF
C/Roc Boronat, 138
08018 Barcelona
sara.rodriguez.fernandez@upf.edu

Roberto Carlini
DTIC, UPF
C/Roc Boronat, 138
08018 Barcelona
roberto.carlini@upf.edu

Leo Wanner
ICREA y DTIC, UPF
C/Roc Boronat, 138
08018 Barcelona
leo.wanner@upf.edu

Resumen: Las combinaciones recurrentes y arbitrarias de palabras (*colocaciones*) son clave para el aprendizaje de lenguas pero presentan dificultades incluso a los estudiantes más avanzados. El uso de herramientas eficientes destinadas al aprendizaje de colocaciones supondría una gran ayuda, sin embargo, las que existen actualmente intentan corregir colocaciones erróneas sin diferenciar entre los distintos tipos de errores ofreciendo, como consecuencia, largas listas de colocaciones de muy diversa naturaleza. Además, sólo se consideran los errores léxicos, dejando de lado los gramaticales que, aunque menos frecuentes, no pueden ignorarse si el objetivo es desarrollar una herramienta capaz de corregir cualquier colocación errónea. En el presente trabajo se propone un método de clasificación automática de errores colocacionales gramaticales cometidos por estudiantes de español estadounidenses, como punto de partida para el diseño de estrategias de corrección específicas para cada tipo de error.
Palabras clave: Aprendizaje de lenguas, colocaciones, tipolgía de errores colocacionales, clasificación de errores gramaticales colocacionales

Abstract: Arbitrary recurrent word combinations (*collocations*) are a key in language learning. However, even advanced students have difficulties when using them. Efficient collocation aiding tools would be of great help. Still, existing "collocation checkers" still struggle to offer corrections to miscollocations. They attempt to correct without making any distinction between the different types of errors, providing, as a consequence, heterogeneous lists of collocations as suggestions. Besides, they focus solely on lexical errors, leaving aside grammatical ones. The former attract more attention, but the latter cannot be ignored either if the goal is to develop a comprehensive collocation aiding tool, able to correct all kinds of miscollocations. We propose an approach to automatically classify grammatical collocation errors made by US learners of Spanish as a starting point for the design of specific correction strategies targeted for each type of error.
Keywords: Second language learning, collocation, collocation error typology, grammatical collocation error classification

1 Introduction

Over the last decades, collocations, i.e., idiosyncratic word co-occurrences such as *spend time*, *take* [*a*] *leave*, *fierce heat*, *deep concern*, and so on have attracted increasing attention of research not only in computational lexicography and lexicology, but also in second language learning (Granger, 1998; Lewis, 2000; Nesselhauf, 2004; Nesselhauf, 2005; Lesniewska, 2006; Alonso Ramos et al., 2010). Studies indicate that collocations are a real challenge for language learners and that they are difficult to master even by advanced students (Nesselhauf, 2003; Bahns and Eldaw, 1993). Wible et al. (2003) show that collocation errors are the most frequent errors found in the writings of students. Orol and Alonso Ramos (2013)' study furthermore reveals that the "collocation density" in learner corpora is nearly the same as in na-

ISSN 1135-5948

tive corpora, i.e., that the use of collocations by learners is as common as it is by native speakers. At the same time, they also find that the collocation error rate in learner corpora is about 32% (compared to about 3% by native speakers). That is, automatic collocation error detection and correction in the context of *Computer Assisted Language Learning* (CALL) could be of great aid to support the learners for better mastering of collocations.

Since the pioneering work by Shei and Pain (2000), several "collocation checkers" have been developed. Most often, these checkers draw upon a collocation list extracted from a reference corpus to compare a collocation used by the student with those in the list (or with variants of those in the list) and thus to detect possible miscollocations (Chang et al., 2008; Park et al., 2008; Östling and Knutsson, 2009; Wu et al., 2010; Dahlmeier and Ng, 2011; Kanashiro Pereira, Manguilimotan, and Matsumoto, 2013) and then potentially offer a list of possible corrections (filtered or ranked according to different metrics).

However, no matter what technique is behind them, state-of-the-art collocation checkers suffer from two main limitations. Firstly, they are able to offer as miscollocation correction suggestions merely large heterogeneous lists of collocations in which one of the words involved in the miscollocation occurs. The learner is thus left with the task of identifying the most appropriate correction by themselves. But this is usually a rather complex task for a language student since selecting a collocation from a list implies that the student knows the meaning of all the collocations in the list, or spends extra time trying to find it. Secondly, they focus only on most common variants of miscollocations.

Both limitations are due to the fact that collocation checkers do not distinguish so far between different types of miscollocations, let alone address all types of miscollocations. Alonso Ramos et al. (2010) argue that collocation errors may be very different in their nature and provide a detailed typology of miscollocations. Comprehensive collocation error type-specific correction techniques would thus most certainly improve the correction performance. However, in order to be able to develop such techniques, we must first be able to classify detected miscollocations, for instance, with respect to Alonso Ramos et al. (2010)'s typlogy. This is the goal of our work.

Alonso Ramos et al. (2010)'s miscollocations typology distinguishes at the first level *grammatical* vs. *lexical* collocation errors. Grammatical collocation errors are more subtle. At the same time, they are also quite common: according to Alonso Ramos et al. (2010), 38% of the miscollocations contain grammatical errors. Therefore, we focus, in what follows, on the automatic classification of grammatical miscollocations.

In the following section, we define in more concrete terms the notion of collocation we use in our work and introduce Alonso Ramos et al. (2010)'s miscollocations typology, which we use in our experiments. Section 3 presents the experiments, and in Section 4, the results of these experiments are discussed. Section 5, finally, outlines the conclusions and our future work in the area of miscollocation classification and correction.

2 Fundamentals on collocations

2.1 The notion of collocation

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; (Choueka, 1988; Church and Hanks, 1989; Pecina, 2008; Evert, 2008; 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 since it seems most useful in the context of second language learning.

2.2 Grammatical miscollocation typology

The typology suggested by Alonso Ramos et al. (2010) groups collocation errors according to three parallel dimensions. The first dimension refers to the location of the error, i.e., whether the collocation as a whole is incorrect or whether one of its elements (the base or the collocate) is incorrect. The second dimension presents differentiations of the characterization of the linguistic phenomena that were observed in miscollocations. The most global differentiation level suggests three error types: lexical, grammatical, and register. The third dimension captures the possible reasons why collocation errors are produced, both interlingual and intralingual. As mentioned above, we focus on the grammatical errors of the second dimension.

Grammatical errors are divided into eight different types:

1. *Determination errors*: Errors resulting from the omission of a determiner when it is required by the collocation, or from its use when the collocation does not accept it; cf., e.g.: **terminar escuela* 'to finish school', where the determiner is expected in Spanish, but is missing.

2. *Number errors*: Errors produced when either the plural or the singular form of a lexical unit is required for a particular collocation, but the opposite is chosen; cf., e.g., **estamos en vacación* 'to be on holiday', where the singular form is used when plural is needed.

3. *Gender errors*: Errors resulting from the choice of the incorrect gender form of the base; cf., e.g., **pasar los vacaciones* 'to spend the holidays'.

4. *Government errors*: Errors produced when the governing preposition of the base or the collocate is missing or mistakenly chosen, or when a preposition is used when there should be none; cf., e.g., **ver a la película* 'to watch a movie, lit. to watch at a movie'. In Spanish, the preposition *a* is required for a direct object only when it refers to people.

5. *Governed errors*: Errors resulting from the wrong use or omission of a preposition that governs the whole collocation; cf., e.g., **estar en buen humor* 'to be in a good mood', instead of *estar de buen humor*.

6. *Specification errors*: Errors produced when a modifier of the base is missing; cf., e.g., **hacer un aterrizaje* 'to make a landing', where the modifier *forzoso* is needed.

7. *Pronoun errors*: Errors resulting from the inappropriate use or the absence of the reflexive pronoun of a verbal collocate; cf., e.g., **las plantas mueren*, 'plants die', where apart from the incorrect lexical choice of the collocate, *morir* instead of *secar*, the reflexive particle *se* is missing.

8. *Order errors*: Errors produced when the base and the collocate appear in the wrong order; cf., e.g., **reputación mala* 'bad reputation', instead of *mala reputación*.

We found that types 5 and 6, i.e. *Governed* and *Specification* errors, are very seldom. For this reason we opted not to consider them at this stage of the experiments.

3 Experiments

The examples of grammatical miscollocation types above illustrate that some of the grammatical error types (e.g., the *Gender* errors and *Order* errors), can be considered a problem of a grammar checker rather than of a collocation checker. We address them nonetheless in the context of collocation verification and correction because they make a collocation to be incorrect.

3.1 Methodology

We developed a set of functions. Each function focuses on the identification of one specific type of grammatical error in given miscollocations. Each function has thus been designed taking into account both the specific particular characteristics of the type of error it deals with and the possibility of a collocation being affected by several errors at the same time, either grammatical, lexical, register or any combination of them. All six functions (recall that we neglect two types of grammatical errors for the moment) receive as input miscollocations found in writings by learners of Spanish, and most of them use a reference native corpus of Spanish (henceforth, RC). In what follows, we briefly describe each one of them.

Determination errors. This function queries the RC to look up common occurrences of both the base and the collocate of the miscollocation, including those with the presence of a determiner and those in which no determiner is found. If the number of occurrences with the determiner is significantly higher than the number of occurrences without the determiner, the collocation is consid-

ered to require a determiner. In this case, if the context of the miscollocation does not contain a determiner, a determination error is flagged. Along the same lines, if it is determined that the collocation does not take a determiner, but the learner uses one, again, a determination error is flagged.

Number errors. Number errors can affect both the base and the collocate and are not necessarily manifested in terms of the lack of concordance, as, e.g., in **tener una vacación* 'to have a holiday', **dimos bienvenidas* 'to welcome', **gané pesos* 'to put on weight', etc. In order to check whether a collocation contains a number error, the corresponding function retrieves from the RC combinations of the lemmas of the base, collocate and the prepositions that depend on the dependent element. In other words, given a preposition, all possible combinations of the forms of the base and the collocate with that particular preposition are retrieved. Then, alternative number forms of the base and collocate are generated (i.e., if an element in the miscollocation is in plural, its singular form is generated, and vice versa) and occurrences of their combinations are retrieved from the RC. If the original form is not one of the possible combinations retrieved from the RC, but any of the alternatives is, the miscollocation is assumed to contain a number error.

Gender errors. Only miscollocations that have a noun as their base can contain this kind of error. However, the form of the base is rarely erroneous (cf., e.g., **pasar los vacaciones*). Rather, there is often a lack of concordance between the base and its determiner, or between the base and the collocate (in N–Adj collocations), resulting from the wrong choice of the gender of the determiner respectively collocate. For this reason, the corresponding function checks the gender of the determiner and adjectives of the base of the given miscollocation. Both the frequency of the miscollocation n-gram (i.e., string consisting of the collocate and the base with its determiner) and linguistic information are considered. For each miscollocation, the function retrieves from the RC the frequency of the original n-gram. Then, it generates new alternatives by changing the gender of the determiner (in VN, NN or prepositional collocations) or the adjective (in NAdj collocations) and looks for the frequency of the new combinations. If this happens to be higher than the frequency of the miscollocation, a gender error is assumed. Otherwise, the concordance between the base and the determiner respectively collocate is checked. If no concordance is found, a gender error is assigned.

Government errors. For identifying this kind of error, we take into account the context in which the miscollocation appears. For this purpose, first, syntactic patterns that contain the miscollocation's base and collocate and any preposition governed by either of the two are retrieved from the RC. Then, it is looked up whether the original syntactic miscollocation pattern that involves a governed preposition appears in the retrieved list. If this is not the case, the miscollocation is assumed to contain a government error.

Pronoun errors. In order to identify pronoun errors, a similar approach to the one used for recognizing determination errors is followed. In this case, frequencies of the combinations with and without reflexive pronouns are retrieved and compared to the miscollocation.

Order errors. To identify an order error, the frequency of the given miscollocation in the RC is calculated. Then, the frequencies of all the possible permutations of the elements of the collocation are compared to the frequency of the miscollocation. If any of them is significantly higher, the collocation is considered to contain an order error.

3.2 Experimental setup

For our experiments, we used a fragment of the Spanish learner corpus CEDEL2 (Lozano, 2009). CEDEL2 is composed of writings of native speakers of US English with different levels of proficiency in Spanish, from 'low-intermediate' to 'advanced'. The writings are of different styles and on different topics (opinion essays, accounts of some past experience, descriptions and letters, etc.). In total, we used 517 texts, with an average of 500 words. Each text was annotated with both correct and incorrect collocations. The number of miscollocations ascended to 1145. Table 1 shows the number of annotated instances for all eight grammatical collocation errors. Our reference corpus consisted of 7 million sentences from newspaper material in Spanish, stored and indexed in Solr. To obtain syntactic dependency information used

in some of the error recognition functions, both corpora were processed with Bohnet (2010)'s dependency parser.

Class	#Instances
Determination errors	146
Number errors	44
Gender errors	77
Government errors	225
Governed errors	2
Specification errors	1
Pronoun errors	28
Order errors	28

Table 1: Number of instances of the grammatical collocation errors annotated in CEDEL2

3.3 Results

Table 2 shows the classification accuracy of the individual grammatical error identification functions for both the positive (collocations containing the type of error that is to be identified) and the negative cases (incorrect collocations affected by any kind of error, except the one that is dealt with).

Type of error	(+)	(−)
Determination	0.719	0.793
Number	0.659	0.851
Gender	0.818	0.989
Government	0.68	0.708
Pronoun	0.357	0.99
Order	0.75	0.848

Table 2: Accuracy of the error detection functions

4 Discussion

We carried out an analysis of the misclassified instances for each experiment, both for the positive and negative classes. In what follows, we present some examples that illustrate the most relevant findings for each type of error.

In all functions, the error identification has been negatively influenced by: (i) the presence of multiple errors in collocations, which causes that queries to the RC do not retrieve any information, and (ii) the automatic preprocessing of the CEDEL2 corpus (note that we are dealing with writings by language learners; the sentences are thus often ungrammatical, such that the error rate of the preprocessing tools (lemmatizer, POS-tagger, morphology-tagger and parser) is considerably higher than in native texts).

Determination errors. As illustrated in the examples (1–2), some determination errors are not identified as such because these collocations can be found both with and without determiner, depending on the context. For instance, a determiner can be required by a specifier, as in (1). Also, we find a singular form of the collocation with a determiner, as in (2), where *tener un hijo* 'to have a child' is correct.

(1) **tiene una reputación*, instead of *tiene reputación* 'to have a reputation'

(2) **tiene los hijos*, instead of *tiene hijos* 'to have children'

With regard to the negative case, i.e., the classification of miscollocations that contain other kinds of errors as determination error, the same reasons can be identified as the source of error. In the following examples, the forms including a determiner, i.e., the singular forms, are more frequent than the forms that do not have it, such that they are classified as determination error.

(3) **tengo planes*, instead of *tengo planes (de)* 'to have plans'

(4) **dijo secretos*, instead of *contar, revelar secretos* 'to tell secrets'

(5) **hacer decisiones*, instead of *tomar decisiones* 'to take decisions'

Number errors. Most failures to identify a number error are due to the fact that, because of multiple errors appearing in the collocation, no usable patterns are retrieved from the RC. A number of failures occur when a collocation is *per se* valid in Spanish, but incorrect in the particular context in which it is used by the learner; cf. (6) and (7).

(6) **fuimos a un museo*, compared to *fuimos a museos* 'to go to museums'

(7) **tienen razón*, compared to *tienen razones* 'to have reasons'

The same occurs in miscollocations that contain other types of errors, but are classified as number error; cf. (8) and (9). An additional source of failure in the negative case is the appearance of lexical errors in the miscollocation, as, e.g., in (10), where a wrong selection of an element of the collocation leads to a correct collocation with a different meaning.

(8) **tener los derechos*, compared to *tener el derecho* 'to have the rights'

(9) **tiene opciones*, compared to *tiene opción* 'to have options'

(10) **hacer divisiones*, compared to *causar divisiones* 'to cause separation', lit. 'to make mathematical divisions'

Gender errors. The analysis of the incorrectly classified instances of both 'gender' and 'other' miscollocations shows that the misclassification is mainly due to errors resulting from the automatic processing of the writings of the students. For instance, in the case of (11–13), the first step of our function returns no information, since all three collocations are affected by several errors and therefore, no valid patterns are retrieved from the RC. To account for this case, concordance is checked. In (11), both the determiner and the base have been assigned masculine gender, so no concordance error was found and the collocation was classified as 'other'. Similarly, *canoa* was incorrectly tagged and no concordance error was found either. Finally, in (13) a parsing error is responsible for the incorrect assignation of the class, since the determiner appears as depending on the verb.

(11) **rechazar los metas*, instead of *alcanzar, lograr las metas* 'to reach goals'

(12) **hacer el canoa*, instead of *ir en canoa* 'canoeing'

(13) **la idioma habla*, instead of *hablar un idioma* 'to speak a language'

As already (11–13), the following 'other error type' miscollocations are affected by several kinds of errors at the same time, which means that concordance has to be checked. Thus, (14) and (15) were incorrectly POS-tagged as N-Adj collocations, such that a concordance between the noun and the adjective was looked for. Since none was found, the collocations were judged to have gender errors.

(14) **sentado por sillas*, instead of *sentado en sillas* 'to sit on chairs'

(15) **completo mis clases*, instead of *termino las clases* 'to complete classes'

Government errors. An analysis of the results for this kind of error reveals that, as already with determination errors, there is often a correct version of the collocation, in this case with a different government, and it is the context which requires the selection of one or the other alternative. Thus, in (16), *tiene el poder* (whithout preposition) should not be used when followed by a verb, but is a possible expression on its own. The same occurs in (17).

(16) **tiene el poder + V*, instead of *tiene el poder (de) + V* 'to have the power (to)' + V

(17) **tener idea + V*, instead of *tener idea (de) + V* 'to have idea (of)' + V

Other types of collocation errors classified as 'government error' are usually caused by lexical errors involved in the collocation, as in the following examples. In (18), a correct collocation can be found with the given base and collocate (*resolución de este problema*). The same can be observed in (19) (*cambiar de religión*). In both cases, there is a correct collocation composed by the original base and collocate and a different preposition, which leads the function to classify them as government errors.

(18) **resolución a este problema*, instead of *solución a este problema* 'solution to a problem'

(19) **cambiar a la religión*, instead of *convertirse a la religión* 'to convert to a religion'

Pronoun errors. The lower accuracy rate for the identification of pronoun errors is due to several reasons. Firstly, due to lexical errors in the same miscollocation, almost a third of the queries to the RC does not retrieve any frequencies. Secondly, lexical errors produce combinations in Spanish that are not necessarily collocations. Thus, *sacar una operación a flote/adelante* (cf. 20) is correct, but it is not a binary collocation. Thirdly, multiple grammatical errors also give place to possible occurrences, as in (21). Finally, there are collocations that accept both the pronominal form and the bare verb form (cf. 22), where it is the context that marks one or the other use.

(20) **sacar una operación*, instead of *hacerse una operación* 'to have surgery'

(21) **aprovecharme de la oportunidad*, instead of *aprovechar la oportunidad* 'to take the most of an opportunity'

(22) **volver loco*, instead of *volverse loco* 'to go mad'

On the contrary, very few collocations of the class 'other error type' have been incorrectly classified as pronoun error. These are cases in which both the pronominal form and the bare verb form are possible, as in (23–24), or where a lexical error gives rise to an acceptable combination (25).

(23) **ir de vacaciones*, compared to *irse de vacaciones* 'to go on holidays'

(24) **cambios producido*, instead of *producirse cambios* 'produced changes'

(25) **darnos la idea*, instead of *hacernos una idea* 'to get an idea'

Order errors. Misclassified order errors are often produced when neither the original combination nor the generated alternatives are found in the RC. As seen before, this is due to multiple errors, such as in (27) and (28). Another source of error, however, can be seen in (26): the use of superlatives, which make the combinations less likely to appear in the RC.

(26) **amigas buenísimas*, instead of *buenísimas amigas* 'close friends'

(27) **nativa parlante*, instead of *hablante nativa* 'native speaker'

(28) **sumamente creo*, instead of *creo firmemente* 'to strongly believe'

As far as other types of errors that are classified as 'order error' are concerned, the most frequent reason is the case of an incorrect collocation, a reordering or appearance in the RC with a higher frequency. Thus, *el día en, buscar trabajo por* and *problemas hacen* in the following examples are acceptable combinations within a sentence.

(29) **en el día*, instead of *durante el día* 'in the day'

(30) **buscar por trabajo*, instead of *buscar trabajo* 'look for jobs'

(31) **hacen problemas*, instead of *causan problemas* 'to cause trouble'

5 Conclusions and future work

Our results show that it is possible to identify grammatical collocation errors in incorrect collocations found in the writings of foreign language learners of Spanish. Most failures to do so are due to following three main reasons: (i) errors during the automatic preprocessing of the learner and reference corpora, (ii) multiple lexical and / or grammatical errors involved in the same collocation, and (iii) valid collocations being grammatically incorrect in the given context. While (i) is not within our reach, (ii) can be partially solved by designing correction strategies that first address lexical errors and attempt to identify grammatical errors only when the lexical correction has been carried out. With respect to (iii), further research will be carried out. Our investigations show that the context in which a collocation appears is essential to identify the type of error involved. Therefore, in the future we plan to explore the use of context in more depth.

6 Acknowledgements

This work has been funded by the Spanish Ministry of Science and Competitiveness (MINECO), through a predoctoral grant with reference BES-2012-057036, in the framework of the project HARenES, under the contract number FFI2011-30219-C02-02.

References

Alonso Ramos, M., L. Wanner, O. Vincze, G. Casamayor, N. Vázquez, E. Mosqueira, and S. 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.

Bahns, J. and M. Eldaw. 1993. Should we teach efl students collocations? *System*, 21(1):101–114.

Bohnet, B. 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.

Bouma, G. 2010. Collocation extraction beyond the independence assumption. In *Proceedings of the ACL 2010, Short paper track*, pages 109–114, Uppsala.

Chang, Y.C., J.S. Chang, H.J. Chen, and H.C. Liou. 2008. An Automatic Collocation Writing Assistant for Taiwanese EFL learners. A case of Corpus Based NLP Technology. *Computer Assisted Language Learning*, 21(3):283–299.

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

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

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

Dahlmeier, D. and H.T. Ng. 2011. Correcting semantic collocation errors with l1-induced paraphrases. In *Proceedings of*

the Conference on Empirical Methods in Natural Language Processing, pages 107–117. Association for Computational Linguistics.

Evert, S. 2008. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin, pages 1212–1248.

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

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

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

Hausmann, F.-J. 1984. Wortschatzlernen ist kollokationslernen. zum lehren und lernen französischer wortwendungen. *Praxis des neusprachlichen Unterrichts*, 31(1):395–406.

Kanashiro Pereira, L.W., E. Manguilimotan, and Y. Matsumoto. 2013. Automated collocation suggestion for japanese second language learners. *ACL 2013*, page 52.

Lesniewska, J. 2006. Collocations and second language use. *Studia LingÃ$\frac{1}{4}$Ãstica Universitatis lagellonicae Cracoviensis*, 123:95–105.

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

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

Mel'čuk, I. 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*. Lawrence Erlbaum Associates, Hillsdale, pages 167–232.

Nesselhauf, N. 2003. The use of collocations by advanced learners of english and some implications for teaching. *Applied linguistics*, 24(2):223–242.

Nesselhauf, N. 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*. Benjamins Academic Publishers, Amsterdam, pages 109–124.

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

Orol, A. and M. Alonso Ramos. 2013. A Comparative Study of Collocations in a Native Corpus and a Learner Corpus of Spanish. *Procedia–Social and Behavioural Sciences*, 96:563–570.

Östling, R. and O. Knutsson. 2009. A corpus-based tool for helping writers with swedish collocations. In *Proceedings of the Workshop on Extracting and Using Constructions in NLP*, pages 28–33.

Park, T., E. Lank, P. Poupart, and M. Terry. 2008. Is the sky pure today? awkchecker: an assistive tool for detecting and correcting collocation errors. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, pages 121–130. ACM.

Pecina, P. 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.

Shei, C.-C. and H. Pain. 2000. An ESL writer's collocational aid. *Computer Assisted Language Learning*, 13(2):167–182.

Wible, D., C.-H. Kuo, N.-L. Tsao, A. Liu, and H.-L. Lin. 2003. Bootstrapping in a language learning environment. *Journal of Computer Assisted Learning*, 19(1):90–102.

Wu, J.-C., Y.-C. Chang, T. Mitamura, and J.S. Chang. 2010. Automatic collocation suggestion in academic writing. In *Proceedings of the ACL Conference, Short paper track*, pages 115–119, Uppsala.