Exploiting a learner corpus for the development of a CALL environment for learning Spanish collocations

Orsolya Vincze, Margarita Alonso Ramos, Estela Mosqueira Suárez, Sabela Prieto González
Universidade da Coruña
Campus da Zapateira, s/n, 15071 A Coruña
E-mail: {ovincze|lxalonso|estela.mosqueira}@udc.es

Abstract

This paper provides an insight into ongoing research focusing on the exploitation of data from learner corpus in order to enhance the performance of an automatic tool aimed at the correction of collocation errors of L2 Spanish speakers. The procedure adopted for collocation annotation is described together with the main difficulties involved in the annotation task, such as the problem of distinguishing collocations from other kinds of idiomatic expressions and from free combinations, the problem of correction judgment, and the problem of assigning concrete error types. It is shown that the fine-grained typology used in the course of error annotation sheds lights on certain collocation error types that are generally not taken into account by automatic error correction tools, such as errors concerning the base of the collocation, target language non-words, and grammatical collocation errors.

Keywords: collocations; learner corpus; error typology; Spanish as second language; Computer assisted language learning (CALL)

1. Introduction

The present paper forms part of a research project that aims at the development of a CALL environment for learning Spanish collocations. The intended CALL environment is conceived of as a flexible and dynamic tool, which will provide an integrated interface combining several resources such as a collocation dictionary (see Alonso Ramos et al., 2010; Vincze et al., 2011[1]), corpora, and an automatic correction tool.

Following Hausmann (1989) and Mel'čuk (1998), we hold that collocations are restricted binary combinations of two lexical units, where one of the two elements, the base, conditions the choice of the other, the collocate. These idiomatic combinations are considered a major challenge for L2 acquisition. In fact, the difference in collocational knowledge has been found to constitute an important factor that contributes to the difference between native and non-native language use (e.g. Howarth, 1998; Granger 1998; Higueras García, 2006).

Previous work suggests that a CALL environment focusing on collocations can profit from data on learners' actual language behavior obtained from corpus research (Shei and Pain, 2000; Chang et al., 2008). In order to gain information on the collocation knowledge and typical errors of Spanish L2 learners, we annotated correct and erroneous collocations in a portion of the CEDEL2 corpus (Lozano, 2009), a corpus containing essays written by English mother tongue Spanish L2 learners.

This paper is structured in the following way. We start, in Section 2, with a brief review of previous work in relation to collocations in the two main research fields concerning our study: learner corpus and correction tools inside a CALL environment. Section 3 provides a description of the collocation annotation procedure we adopted, and looks into the three main difficulties involved in the task: 1. recognition of collocations, 2. correction judgment, and 3. error type annotation. Following this, in Section 4, we highlight some characteristics of collocation errors observed in the corpus that generally have not been attended to by automatic correction tools; these are: 1. the location of the error, 2. L1 interference and 3. grammatical collocation errors. Finally, in Section 5, we draw some conclusions on the presented work and give future lines of research.

2. Previous work

As we have mentioned, the object of the present study lies at the crossroads of two fields of research. Within the first of these, learner corpus research, collocations have been the subject of a considerable number of studies since Granger's (1998) seminal work. These constitute quantitative and qualitative studies comparing learners' and native speakers' collocation production. However, none of them goes into such detail in terms of error analysis as our own research. Some of the most recent studies in this field are Nesselhauf (2005), Martelli (2007) and Thewissen (2008).

Similarly, in the field of CALL, we find several proposals aiming at creating an automatic tool for the correction of collocation errors. Notably, some of these, such as Shei and Pain (2000), Liu (2002) and Chang et al. (2008), make use of error analysis data coming from learner corpora. The first proposal uses corpus data to build an error library to enhance the performance of the system, while Liu (2002), in addition to this, also exploits her observation that most erroneous collocates can be related to their correct counterparts through

[1] Diccionario de Colocaciones del Español, available at: www.dicesp.com

semantic relations established in WordNet (Fellbaum, 1998). Finally, re-examining Liu's (2002) data, Chang et al. (2008) emphasize that the great majority of learner collocation errors can be accounted for by L1 interference. Therefore, their system uses bilingual dictionaries to check synonymous translations of erroneous collocates in order to suggest likely corrections.

3. Corpus annotation

With the aim of studying the collocation production of L2 Spanish learners, we manually annotated 100 essays from the CEDEL2 corpus, amounting to 46420 words. The corpus annotation task was carried out by native speakers of Spanish, following a well-defined procedure, and making use of an elaborate typology devised by Alonso Ramos et al. (2010, see below) for labeling collocation errors. However, as the low inter-annotator agreement shows, the annotation task posed a significant challenge, mainly due to the notoriously fuzzy interpretation of the notion of collocation.

3.1 The annotation process

The corpus annotation was carried out by two main annotators, whose annotations were merged and revised by a third, consensus annotator. Annotations lacking consensus were resolved in the following way: Cases of collocations judged to be correct by one and incorrect by the other main annotator were checked against corpus data by the consensus annotator. When at least five cases of the given collocation were found in the *Corpus de Referencia del Español Actual* (CREA), it was considered a correct combination. Cases that could not be resolved using this method were sent to three independent annotators, and subsequently treated according to the majority vote. Finally, dubious annotations and conclusions on merged annotations were discussed in weekly annotation sessions supervised by an expert annotator.

As for the success of the annotation process, we should note that despite the well-defined annotation procedure and the weekly sessions to comment on criteria concerning the new annotations, we were able to achieve only a slight increase in inter-annotator agreement, which however remained considerably low throughout the whole of the annotation process: an average of about 30% during the first weeks and average of about 50% over the last weeks.

This issue is mainly related to the problem of recognizing collocations in the learner texts. In what follows we discuss this and other difficulties affecting the annotation process, such as the problems of correction judgment, and the problems of interpreting errors.

3.2 Problems of recognizing collocations

The problem of recognizing collocations in the learner texts can be ascribed to the difficulty of establishing clear and, most importantly, operational criteria for delimiting the notion of collocation. In practice, this results in the annotators having difficulty in telling collocations apart from free combinations, on the one hand, and from idioms, on the other hand.

For instance, it is quite straightforward to agree on that *buena nota* 'good grade' is a collocation, given that the semantic characteristics of a noun like *nota* 'grade' call for a qualification adjective. This is not so in the case of the combination *buena comida* 'good food', where the meaning of the noun *comida* 'food' does not necessarily require qualification. Consider, however, the combination *comida rica* 'delicious food', where the adjective, *rico* 'delicious', has a rather restricted use; it is the adjective prototypically chosen to speak about good food. From our point of view, combinations such as *comida rica* should be considered collocations, and, consequently, other less idiomatic combinations, containing less restricted adjectives appearing with the same noun, such as *buena comida* 'good food' or even *comida fantástica* 'fantastic food' will be considered collocations as well. An example for the difficulty of distinguishing collocations from idioms is the case of *darse cuenta* 'realize', which should be treated as a non-compositional expression, given its frozen syntactic structure. It was mistaken for a collocation by the annotators due to the fact that the verb *dar* 'give' is often used in light verb constructions, as in *dar un paseo* 'take a walk', *dar consejos* 'give advice', etc.

We also noticed that correct collocations often passed unnoticed by annotators until an incorrect counterpart of the same combination was found. An example for this is the case of *país de origen* 'country of origin', which was not annotated as a collocation until the erroneous combination *países maternos* lit. 'mother(ly) countries' was found in the corpus. At the same time, any error was bound to be perceived as a collocation error by the annotators. For instance, the free combinations *lleno *con historia* lit. 'full with history' and *recorrimos *por la isla* 'we travelled all over the island' were both annotated in the first stage of the annotation process, probably because the preposition errors made them more salient[2].

3.3 Problems of correction judgment

The main issues here are the individual permissiveness

[2] Note that we do not treat as collocations word combinations consisting of a lexical element and its governed preposition (e.g. *depende de* 'depend on'), often referred to in the literature as *grammatical collocations* (Cf. Benson et al., 1986). However, prepositions governed by a member of a collocation (e.g. *tener miedo de* lit. 'have fear of') are considered to form part of the expression as a whole, therefore, when erroneous, they are annotated as grammatical collocation errors (see below, in Section 4.2).

of annotators, on the one hand, and the challenge posed by language variation, on the other hand. Differences in the individual permissiveness of each annotator towards unusual language use led to lack of consensus in judging a lexical combination correct or incorrect. It also appears that annotators tend to be less permissive with non-native speakers in terms of creative or unusual language use than they would be with native peers.

The problem of language variation was noticed especially in the case of collocations typically used in Latin American Spanish. They were judged, at first sight, as incorrect by the annotators, however corpus data showed that these combinations are actually in use in other Spanish-speaking countries. Consequently, these expressions were annotated in the corpus as correct collocations, specifying the language variant they belong to. For example, the combination *hice las reservaciones* 'I made the reservations' was perceived as incorrect, given that European Spanish uses the form *reserva* 'reservation' and not *reservaciones.* We also find differences, for instance, in the use of collocate verbs such as in *tomar clases* lit. 'take classes' used in America (see in (1)) and the combination expected by European speakers of Spanish in the same context: *ir a clase* lit. 'go to class'.

(1) Empecé *tomando clases* de una española
lit. I started by *taking classes* from a Spanish women

Finally, the limitations inherent to written text, such as missing intonation pattern, sometimes also caused difficulties in the interpretation of the text itself.

3.4 Problems of interpreting errors

Three kinds of problems constituted a challenge when labeling errors with the specific error categories. Firstly, given that the error type labels, to some extent, reflect how the erroneous expression relates to its correction, cases when more than one correction was possible resulted problematic. For instance, in the sentence in (2), the combination *hizo gorditas* can be corrected either for a collocation *ponerse gordas* lit. 'put one selves fat' or a single verb *engordar* 'gain weight'. In the first case the error should be described as the use of an incorrect collocate (*hacer* instead of *ponerse*), while in the second case, it should be described as the use of an erroneous analytical form (*hacer gorditas*) instead of a single lexical item (*engordar*).

(2) el viaje no **nos hizo gorditas*
lit. the trip didn't *make us fatty*
we didn't *gain weight* during the trip

Secondly, some incorrect collocation-like combinations produced by the learners turned out to be literal translations of combinations in the native language that have no collocation equivalent in Spanish. For instance, the erroneous form **humo de segunda mano* corresponds to the English collocation *secondhand smoke*, which can only be translated to Spanish by a complex phrase expressing the same meaning without constituting a phraseological expression: *humo del tabaco de otras personas* 'smoke from other people's cigarette'. On the contrary, some expressions used by the learners do not constitute collocations themselves; however the correct form to be used should be a collocation in Spanish. An example for this case can be seen in (3) where the expression using the copulative verb and the adjective *curioso* 'curious' should be corrected as a collocation: *tengo curiosidad* lit. 'I have curiosity'.

(3) **estoy curiosa* conocerlo
lit. *I'm curious* to get to know it

Thirdly, two coexisting category labels had to be allowed in the cases where the source of the error could not be determined unambiguously. For instance, in the case of the incorrect collocation **hice citas* lit. 'I made appointments', the annotators found it feasible to treat the error both as a direct translation from English and as a generalization error, whereby the generic verb *hacer* 'make/do' is used instead of the correct and more restricted *concertar* 'arrange'.

4. Exploiting corpus data for a learning tool

We have already mentioned that the error typology (Alonso et al. 2010) we used in the annotation task allows for a more detailed error annotation than the coarse-grained typologies used in other learning tools focusing on collocations (Chang et al. 2008; Shei and Pain 2000). In these, only lexical errors affecting the collocate are taken into account, and the main type of error foreseen is that resulting from L1 lexical transfer. With these limitations, a learning tool aimed at the automatic recognition and correction of collocation errors would have difficulties in identifying some of the error types inherent in our typology. In what follows, we will show some particular features revealed by our detailed error analysis.

4.1 The collocation error typology

Our error typology distinguishes three parallel dimensions. The first, "location" dimension captures whether the error concerns one of the elements of the collocation (the *base* or the *collocate*, following Hausmann's (1989) terminology) or the collocation as a whole. The second dimension models *descriptive* error analysis and distinguishes between three main types of error: lexical, grammatical and register error, the first two of which are further detailed in several subtypes. Finally, the third dimension represents *explanatory* error analysis: it concerns the source of the error, described by the main categories of transfer errors, that is, errors reflecting L1 interference and interlanguage errors, resulting from the incomplete knowledge of the L2 without L1 interference.

4.2 The "location" of errors

In contrast with the general approach in automatic collocation error correction, we have taken into account any error affecting either member of the collocation or the expression as a whole, as captured by the "location" dimension of our typology. As a result, we have annotated not only erroneous collocates (4), but also erroneous bases (5). The latter case would pose a difficulty for systems that correct collocation errors merely verifying the correctness of the collocate. For example, in the case of the collocation in (4), where the collocate is incorrect, a search for collocate verbs of the base *regla* 'rule', similarly to Liu's (2002) or Chang et al.'s (2008) proposal, restricted to those synonymous or sharing translation synonyms with *interrumpir* 'to interrupt', would likely return the correct combination. However, when it is the base that is erroneous, such as in (5), the same strategy, a search for co-occurring verbs with the base *gol* 'goal (in sport)' will not be effective. Note that, we have also found cases where both the collocate and the base are incorrect, as in (6).

(4) **interrumpir una regla* 'interrupt a rule' instead of *romper una regla* 'break a rule'
(5) **lograr un gol* 'achieve a goal (in sport)' instead of *lograr un objetivo* 'achieve an aim'
(6) *pasar un testemuño 'pass a testimony (from Portuguese)' instead of *dar testimonio* 'give testimony'

There is a total number of 445 erroneous collocations among the 1401 collocations annotated in the corpus. For now, we will limit our analysis to those affected by lexical errors, a total number of 266 collocations. As for the "localization" dimension, we find lexical errors of the collocate in the highest number, affecting a total of 174 collocations (61%), however a still large proportion, 61 collocations (21%) have erroneous bases, while 50 expressions (18%) contain a lexical error that is considered to affect the collocation as a whole. These numbers suggest that a CALL system aimed at correcting collocation errors efficiently, shouldn't be limited to collocate errors, but should also foresee lexical errors concerning the base of the collocation.

Lexical errors affecting the collocation as a whole are of various kinds: an otherwise correct combination can be used in an incorrect sense, as in (7) where, in order to express the correct meaning, the combination *aliviar el estrés* 'ease the stress' should be substituted for *aumentar el estrés* 'increase the stress'.

(7) al oirlo hablar, tengo que apagar al aparato para no **aliviar el estrés*
when I hear him speak, I have to turn off the television in order to not *to ease the stress*

Furthermore, we also considered here incorrect collocation-like expressions that should be correctly expressed by a single word (8), or, as we have seen above, by a non-idiomatic expression (9) and cases of incorrect single-word forms standing instead of a collocation (10).

(8) **poner apasionado* 'make passionate' instead of *apasionar* 'to fascinate'
(9) **humo de segunda mano* 'secondhand smoke' instead of *humo del tabaco de otras personas* 'smoke from other people's cigarette'
(10) **misenterpretación* 'misinterpretation' instead of *mala interpretación*

The correction of these kinds of expressions may pose further difficulties for an automatic tool.

4.3 L1 influence

Out of the 284 lexical collocation errors found in the corpus (note that a collocation can contain more than one error), 67% were found to be transfer errors, while 33% were annotated as interlanguage errors. This is in line with the findings of other authors such as Liu (2002), Nesselhauf (2005) etc. Our corpus data also corroborates the hypothesis that automatic tools such as Liu (2002), Chang et al. (2008) and Futagi (2010) make use of, that is, in most lexical collocation errors, the erroneous element can be conceived of as a synonym or a translation synonym of its correct counterpart for correction purposes. Remarkably, we find this is true both in the case of L1 transfer and interlanguage errors. Nevertheless, we would like to highlight a few error types that do not fit into this picture.

In the case of L1 transfer errors, we found the example shown in (11). We assume that the word *colegio* 'primary school' is used instead of *universidad* 'university' due to its formal resemblance to the English word *college*. This case shows that errors resulting from the phenomenon commonly known by language learners and teachers as 'false friends', that is the confusion of formally similar but semantically not necessarily related word forms, might be taken into account in language tools.

(11) Hemos **licenciado en el colegio* en la vecina ciudad
Lit. We *earned a degree in the primary school* in the neighbor town

Other phenomena concern the use of lexical elements that constitute non-words in the target language. Firstly, a small group of transfer errors (amounting to less than 6%) involve the use of a L1 lexical item, as in (12), or a lexical item from a L2 different from the target language (TL), see example (6) above. These forms are sometimes adapted to TL orthography and morphology as in (13), where the erroneous form *misenterpretaciones* stands instead of the Spanish collocation *malas interpretaciones* lit. 'wrong interpretations'. Secondly, among interlanguage errors we find cases of Spanish non-words, we assume to be the result of an erroneous derivation

process. For instance, in (14) a non existent wordform **frescar* is derived instead of *refrescarse* 'cool down'.

(12) En Oaxaca se puede **ir de hiking*
Lit. In Oaxaca one can *go hiking*

(13) el trama del libro es una sarta de **misenterpretaciones*
Lit. the plot of the book is string of *misinterpretations*

(14) Las *temperaturas* cambian y **frescan* un poco
Lit. The *temperatures* change and *cool down* a bit

4.4 Grammatical errors

Learner tools aimed at the correction of collocations in general do not take grammatical errors into account at all. An exception to this is Futagi (2010) where article and inflection errors are considered, although merely with the aim of enhancing the performance of collocation extraction from learner texts. Our approach is clearly different from this, given that, from our point of view, certain grammatical errors should be considered as proper collocation errors, due to the fact that they affect the correct formulation of the lexical combination.

Grammatical collocation errors are rather frequent in the corpus, they concern 212 (44%) of the 478 erroneous collocations annotated. In what follows we show examples for each class of grammatical collocation error:

-determination error: **tomar sol* instead of *tomar* el *sol* 'to sunbathe'

-incorrect government: **montar a bicicleta* instead of *montar en bicicleta* 'to ride a bike'

-incorrect gender: **mente abierto* instead of *mente abierta* 'open mind'

-incorrect number: **estamos en vacación* instead of *estamos de vacaciones* 'we are on holiday'

-incorrect external government: **en buen humor* instead of *de buen humor* 'of good humor'

-pronominal verb error: **muero de ganas* instead of *me muero de ganas* lit. 'I am dying from desire [to do something]'

-word order error: **reputacion mala* instead of *mala reputación* 'bad reputation'

5. Conclusions and future work

The present paper has provided an insight into ongoing research focusing on the exploitation of data from learner corpus in order to enhance the performance of an automatic tool aimed at the correction of collocation errors of L2 Spanish speakers.

As we have shown, collocation annotation in corpus is not a straightforward process; the difficulties discussed in more detail are the problem of telling collocations apart from other kinds of idiomatic expressions or from free combinations, the problem of correction judgment, and the problem of assigning concrete error types. We have also demonstrated that the fine-grained typology we used for error annotations sheds lights on certain error types that are generally not taken into account by automatic error correction tools, such as errors concerning the base of the collocation, target language non-words, and grammatical collocation errors.

As for future investigation, our goal is to annotate a comparable corpus of native speakers of Spanish in order to compare the collocation knowledge and use of the native and non-native groups. We also plan to exploit our data on typical collocation errors for automatically generating activities for practicing collocations.

6. Acknowledgements

The work described in this paper has been carried out in the framework of the project COLOCATE, partially funded under the contract number FFI2008-06479-C02-01 by the Spanish Ministry of Science and Innovation (MICINN) and the FEDER Funds of the European Commission.

7. References

Alonso Ramos, M., Nishikawa, A. & Vincze, O. (2010). DiCE in the web: An online Spanish collocation dictionary. In S. Granger, M. Paquot (eds.) *Elexicography in the 21st Century: New challenges, new applications. Proceedings of eLex 2009*. Cahiers du cental 7. Louvain-la Neuve: Presses Universitaires de Louvain, pp. 369-374.

Alonso Ramos, M., Wanner, L., Vincze, O., Casamayor, G., Vázquez, N., Mosqueira, E. & Prieto, S. (2010). Towards a Motivated Annotation Schema of Collocation Erros in Learner Corpora. In N. Calzolari et al. (eds.) *Proceedings of the Seventh conference on International Language Resources and Evaluation* (LREC'10). Paris: ELRA, pp. 3209-3214.

Benson, M., Benson, E.E. & Ilson R. (1986). *The BBI combinatory dictionary of English*. Amsterdam/Philadelphia: John Benjamins.

Chang, Y., Chang, J., Chen, H. & Liou, H. (2008). An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning*, 21(3), pp. 283–299.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Futagi, Y. (2010). The effects of learner errors on the development of a collocation detection tool. In *AND'10 Proceedings of the fourth workshop on Analytics for noisy unstructured text data.* New York: ACM, pp. 27-34.

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.

Hausmann, F.J. (1989). Le dictionnaire de collocations. In F.J. Hausmann *et al.* (eds.) *Wörterbücher – Dictionaries – Dictionnaires*, vol. 1. Berlin: de

Gruyter, pp. 1010-1019.

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

Howarth, P. (1998). The phraseology of learners' academic writing. In A.P. Cowie (ed.) *Phraseology. Theory, Analysis, and Applications*. Oxford: Oxford University Press, pp. 161-186.

Liu, L.E. (2002). *A corpus-based lexical semantic investigation of verb-noun miscollocations in Taiwan learners' English*. Master's thesis, Tamkang University, Taipei.

Lozano, C**.** (2009). CEDEL2: Corpus Escrito del Español L2. In C.M. Bretones Callejas et al. (eds.) *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente.* Almería: Universidad de Almería, pp. 80-93.

Martelli, A. (2007). *Lexical Collocations in Learner English: A Corpus-Based Approach.* Alessandria: Edizioni dell' Orso.

Mel'čuk, I. (1998). Collocations and Lexical Functions. In A.P. Cowie (ed.) *Phraseology. Theory, Analysis, and Applications*. Oxford: Oxford University Press, pp. 23-53.

Nesselhauf, N. (2005). *Collocations in a learner corpus.* Amsterdam/Philadelphia: John Benjamins.

Real Academia Española. *Corpus de referencia del español actual.* Accessed at: http://www.rae.es.

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

Thewissen, J. (2008). The phraseological errors of French-, German- and Spanish-speaking EFL learners: evidence from an error-tagged learner corpus. In *Proceedings from the 8th Teaching and Language Corpora Conference (TaLC8)*. Lisbon: Associação de Estudos e de Investigação Científica do ISLA-Lisboa, pp. 300-306.

Vincze, O., Mosqueira E. & Alonso Ramos, M. (2011). An online collocation dictionary of Spanish. In I. Boguslavsky, L. Wanner, (eds.) *Proceedings of the 5th International Conference on Meaning-Text Theory*, pp. 275-286. Available at: http://meaningtext.net/mtt2011.