GRANGER, S.; PAQUOT, M. (EDS.). 2010. ELEXICOGRAPHY IN THE 21ST CENTURY: NEW CHALLENGES, NEW APPLICATIONS. PROCEEDINGS OF ELEX2009, LOUVAIN-LA-NEUVE, 22-24 OCTOBER 2009. CAHIERS DU CENTAL 7. LOUVAIN-LA-NEUVE, PRESSES UNIVERSITAIRES DE LOUVAIN, 675-380.

Tagging Collocations for Learners

Margarita Alonso Ramos[1], Leo Wanner[2], Nancy Vázquez Veiga[1], Orsolya Vincze[1], Estela Mosqueira Suárez[1], Sabela Prieto González[1]
[1]University of A Coruña, [2]ICREA and Pompeu Fabra University

Abstract

Collocations play a significant role in second language acquisition. In order to be able to offer efficient support to learners, an NLP-based CALL environment for learning collocations should be based on a representative collocation error annotated learner corpus. We are currently working on such a corpus for Spanish, starting from a fine-grained typology of collocation errors and drawing upon an existing learner corpus, namely CEDEL2 from the Autonomous University of Madrid. In this paper, we present this typology and discuss the first findings obtained from our annotation work.

Keywords: collocation, learner corpus, error typology, Spanish as second language

1. Introduction

The importance of *collocations* in second language acquisition is increasingly recognized in the community (Lewis 2000; Granger 1998b; Howarth 1998; Nesselhauf 2003, 2005; Alonso Ramos 2006; Higueras 2006; Martelli 2006). To adequately support students in learning collocations, it is crucial to identify and classify the collocation errors made by them and then offer targeted exercises and adequate illustrative material. This presupposes the availability of collocation tagged learner and general corpora: a learner corpus allows us to identify the most common collocation errors; a general corpus is needed as a source of illustration and training material.

We aim at the development of an advanced NLP-based computer assisted language learning (CALL) environment for learning collocations in Spanish. In this paper, we focus on the problem of processing Spanish learner corpora, which consists of three stages: (i) analysis of the corpus and derivation of a collocation error typology; (ii) definition of a tag set to annotate the corpus; and (iii) tagging the corpus.

2. Towards a collocation error typology

[1] University of A Coruña, lxalonso@udc.es

[2] ICREA and Pompeu Fabra University, leo.wanner@upf.edu. This paper was written within the framework of a research project: FFI2008-06479-C02-01 (Ministerio de Ciencia and partially FEDER).

A detailed analysis of learner corpora has proved to be essential (Dagneaux *et al.* 1998; Granger 1998a, 2007; Tono 2003). Such an analysis requires a predefined error tag set or error typology (Granger 2007). This is also true for the analysis of a collocation learner corpus. Currently available general learner error typologies tend to group collocation errors into a single subclass of lexical errors (Aldabe *et al.* 2005; Milićević and Hamel; 2007; Granger 2007; Díaz-Negrillo and García-Cumbreras 2007). Occasionally, collocation errors are also discussed referring to the POS of the collocation elements (Philip 2007). A closer look at a learner corpus reveals, however that a more detailed typology is needed. For the purpose of the present study, we used the *Corpus Escrito del Español L2* (CEDEL2) from the Autonomous University of Madrid[3], which consists of short compositions written by native speakers of English (L1). Consider some examples from CEDEL2:

(1) *deseo* lograr el gol *de ser bilingual*, lit. 'I desire achieve the goal of being bilingual'

(2) […] llenar un puesto [*de trabajo*], lit. 'fill a position [of work]'

(3) recibí un llamo *de Brad*, lit. 'I received a call from Brad'.

(4) *Algunos* tienen prejuicio *por edad*, lit. 'Some have prejudice for age'

Apart from errors not related to collocations (*e.g. bilingual* instead of *bilingüe*), which we ignore, the following collocation construction errors stand out[4]:

(1′) error in the base resulting from the projection of a word in L1 (English) to L2 (Spanish), as e.g. *goal* → *gol*: *lograr* [*el*] <u>*gol*</u> – instead of *lograr* [*el*] <u>*objetivo*</u>;

(2′) error in the collocate resulting from a literal translation of a word from L1 to L2, as e.g. *fill* → *llenar*: <u>*llenar*</u> [*un*] *puesto* – instead of <u>*ocupar*</u> [*un*] *puesto*;

(3′) error in the base resulting from a wrong morphological derivation and an inappropriate use of the collocation as a whole in the given context, as e.g. *llamar* → *llamo*: *recibí un* <u>*llamo*</u> *de Brad* – instead of *recibí una* <u>*llamada*</u> *de Brad*; or, better: *me llamó Brad*;

(4′) error in the number of the base and in the governed preposition, as e.g. *prejuicio*: *tienen prejuici*<u>*o*</u> [<u>*por*</u> *algo*], instead of *tienen prejuicio*<u>*s*</u> [<u>*hacia*</u> *algo*].

[3] CEDEL2, which has been compiled by the group directed by Amaya Mendikoetxea, contains about 400,000 words of essays written in Spanish by native speakers of English. The essays are classified with respect to the proficiency level of the authors. The essays underlying our study were written by learners with intermediate or advanced level of Spanish. For more information, see http://www.uam.es/proyectosinv/woslac/cedel2.htm.

[4] We interpret collocations in the sense of Hausmann (1979) as idiosyncratic word co-occurrences consisting of a base and a collocate.

The errors are very different. Therefore, a fine-grained collocation error typology is needed to capture these differences and be able to offer adequate didactic means to address them.

In the present stage of our work, we distinguish three main types of collocation errors: lexical errors, grammatical errors and register errors. Lexical errors concern either the whole collocation or one of its elements. In the first case, we find inexistent collocations in Spanish whose meaning would be correctly expressed by a single lexical unit (LU) (*e.g.* **hacer de cotilleos*, lit '[to] make of gossip' instead of *cotillear* '[to] gossip'), and inexistent single LUs used instead of collocations (*e.g.* **escaparatar* instead of *ir de escaparates*, lit. '[to] go of shop window'). In the second case, we distinguish between errors concerning paradigmatic lexical selection (*e.g.* **lograr un gol* lit '[to] achieve a goal (in football)' instead of *lograr un objetivo,* lit '[to] achieve a goal') and errors concerning syntagmatic lexical selection (*e.g.* **escribir el examen,* lit '[to] write the exam' instead of *hacer el examen,* lit '[to] do the exam); the former concern the base, the second the collocate.

Most lexical errors are literal translations from L1. Although a finer distinction is necessary later on to determine the source of errors, as a first approximation, the distinction between "transfer by importation", i.e., adoption of an inexistent form in L2 – *recibir un llamo*, lit '[to] receive a call', instead of *recibir una llamada* – and "transfer by extension", i.e., extension of the meaning of an L2 lexical unit – *salvar dinero*, lit. '[to] save money', instead of *ahorrar dinero* – is valid.

Grammatical errors in our typology are directly linked to collocations. They concern information that a learner cannot derive from the grammar of L2 and that must be described in the entry for the base of the collocation (*e.g.* **hablar al teléfono* lit. '[to] speak to the phone' instead of *hablar por teléfono* '[to] speak through the phone').

In the class of register error, we group collocations that are pragmatically inappropriate. Thus, *tengo el deseo de ser bilingüe*, lit. 'I have the desire of being bilingual' sounds odd in an informal context – better: *me gustaría ser bilingüe* 'I would like to be bilingual'.

3. The process of tagging collocations in CEDEL2

Apart from a collocation error typology, a detailed semantic typology of collocations is crucial in order to be able to offer the learner examples of similar collocations. The most detailed and systematic semantically-oriented typology of collocations we know of are the Lexical Functions (Mel'čuk, 1996), from now on referred to as LFs.

With the collocation error and the LF typologies at hand, we tag all collocations in CEDEL2. In the case of collocation errors, we also annotate the correct version of the erroneous collocation and the corresponding LF. Consider the following examples.

(1´´) *lograr* [*el*] *gol*: lexical error in the base; extension of the meaning of Sp. *gol* ‘goal (in football)’ due to phonetic similarity with Eng. *goal*; LF: Real1; correct: *lograr* [*el*] *objetivo*

(2´´) *llenar* [*un*] *puesto*: lexical error in the collocate; extension of the meaning of Sp. *llenar* ‘fill’ based on the English collocation [*to*] *fill a position*; LF: Oper1; correct: ocupar [un] puesto

(3´´) *recibí un llamo* [*de Brad*]: lexical error in the base; erroneous derivation based on the first person singular form of the verb Sp. *llamar*, possibly analogous with forms like *paseo<pasear*, *canto<cantar*, etc.; LF: Oper2; correct: *recibí una llamada* [*de Brad*]

Example (4’’) shows that, on the one hand, a single collocation may show more than one error, and, on the other hand, that the determination of the source of an error is not always straightforward.

(4´´) *tienen prejuicio* [*por algo*]: 1. grammatical error in the government of the base; intralingual; 2. grammatical error in the number of the base; intralingual or possibly interlingual since Eng. *prejudice* can be used both as a countable or an uncountable noun; LF: Oper1; correct: *tienen prejuicios* [hacia algo].

The tagging of the learner corpus is currently being performed manually, supported by an interactive annotation tool, *Knowtator*, which is realized as a plug-in of the knowledge acquisition framework Protégé. The application allows us to define an annotation schema used in the process of annotation to give information on the semantics of the combinations– through LFs –, and, in the case of erroneous collocations, to describe the errors and propose a correction. We are also about to develop a collocation tagger that will tag both LFs and collocation errors. The work on the LF-tagger draws upon the work described in Wanner *et al.* (2006).

4. Conclusion

The preliminary evaluation of the corpus we annotated so far in accordance with the schema presented above, reveals that 39% of the collocations used by learners contain some error. 62% of the erroneous collocations contain lexical errors, 33% show grammatical errors, whereas 5% have both lexical and grammatical errors. In a more fine-grained analysis of the more prominent lexical errors, we find that 54% of these represent an incorrect choice of the collocate, 20% the use of an incorrect base, 16% the use of an existing collocation with a different sense, while 10% are cases of using collocation-type constructions instead of single LUs. As for the possible source of errors, we can establish that the great majority of lexical errors – 70% – represent clear cases of lexical transfer from L1 to L2. However, further investigation based on a larger annotated corpus is needed to draw more fine-grained conclusions. We are thus currently working on the extension of our collocation error annotated learner corpus.

References

ALDABE, I., ARRIETA, B., DÍAZ DE ILARRAZA, A., MARITXALAR, M., ORONOZ, M. and URIA, L. (2005). Propuesta de una clasificación general y dinámica para la definición de errores. *Revista de Psicodidáctica*, 10/2: 47-60.

ALONSO RAMOS, M. (2006). Towards a dynamic way of learning collocations in a second language. In E. Corino, C. Marello and C. Onesti (eds) *Proceedings XII EURALEX International Congress*, Torino, Italy, September 6th-9th 2006. Alessandria: Edizioni Dell'Orso: 909-921.

DAGNEAUX, E., DENNESS, S. and GRANGER, S. (1998). Computer-aided error analysis. *System*, 26: 163-174.

DÍAZ-NEGRILLO, A. and GARCÍA-CUMBRERAS, M.A. (2007). A tagging tool for error analysis on learner corpora. *ICAME Journal*, 31/1: 197-203.

GRANGER, S. (ed.) (1998a). *Learner English on Computer*. Oxford: Oxford University Press.

GRANGER, S. (1998b). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (ed.) *Phraseology. Theory, Analysis, and Applications*. Oxford: Clarendon Press : 145-160.

GRANGER, S. (2007). Corpus d'apprenants, annotation d'erreurs et ALAO: une synergie prometteuse, *Cahiers de lexicologie*, 91/2 : 465-480.

HAUSMANN, F. J. (1979). Un dictionnaire des collocations est-il possible? *Travaux de littérature et de linguistique de l'Université de Strasbourg*, 17/1: 187-195.

HIGUERAS, 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: Clarendon Press: 161-186.

LEWIS, M. (2000). *Teaching collocation. Further developments in the lexical approach.* London: Language Teaching Publications.

MARTELLI, A. (2006). A corpus-based description of English lexical collocations used by Italian advanced learners. In E. Corino, C. Marello and C. Onesti (eds) *Proceedings XII EURALEX International Congress*, Torino, Italy, September 6th-9th 2006. Alessandria: Edizioni Dell'Orso: 1005-1012.

MEL'ČUK, I. (1996). Lexical Functions: A tool for the Description of Lexical Relations in the Lexicon. In L. Wanner (ed.). *Lexical functions in lexicography and natural language processing*. Amsterdam and Philadelphia: John Benjamins: 37-102.

MILIĆEVIĆ, J. and M-J., HAMEL. (2007). Un dictionnaire de reformulation pour les apprenants du français langue seconde. In G. Chevalier, K. Gauvin and D. Merkle (eds) *Actes du 29^e^ Colloque annuel de l'ALPA tenu a l'Université de Moncton*, Moncton, Canada, November 4th-5th 2005. *Revue de l'Université de Moncton,* numéro hors série: 145-167.

PHILIP, G. (2007). Decomposition and delexicalisation in learners' collocational (mis)behaviour. In M. Davies, P. Rayson, S. Hunston and P. Danielsson (eds) *Online Proceedings of Corpus Linguistics 2007*, Birmingham, United Kingdom, July 27th-30th 2007. Birmingham: University of Birmingham: 1-11.

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. (2005). *Collocations in a learner corpus.* Amsterdam and Philadelphia: John Benjamins.

NESSELHAUF, N. and TSCHICHOLD, C. (2002). Collocations in CALL: An investigation of vocabulary-building software for EFL, *Computer Assisted Language Learning,* 15/3: 251-279.

TONO, Y. (2003). Learner corpora: Design, development and applications. In D. Archer *et al*. (eds.) *Proceedings of the Corpus Linguistics 2003*, Lancaster, United Kingdom, March 28th-31th 2003. Lancaster: Lancaster University, University Centre for Computer Corpus Research on Language: 323-343.

WANNER, L., BOHNET, B. and GIERETH, M. (2006). Making sense of collocations. *Computer Speech & Language*, 20/4: 609-624.