Deriving criteria for teaching Spanish collocations from a learner corpus

Marcos García Salido, Margarita Alonso Ramos and Orsolya Vincze
Universidade da Coruña
marcos.garcias@udc.es

The aim of this paper is to explore the possibilities offered by learner corpus analysis for deriving criteria that help us determine what types of collocations are especially problematic for SFL learners and, therefore, should be paid special attention in teaching.

The study uses data extracted from sections of the CEDEL2 corpus (Lozano and Mendikoetxea, 2013) in which collocations were manually annotated by at least two native speakers (see Alonso Ramos et al., 2010a). We adopted the definition established within the phraseological tradition (Hausmann, 1989; Mel'čuk, 1996), according to which collocations are constituted by two elements, the *base*, freely chosen by the speaker and the *collocate*, whose selection is restricted by the base. Note that this definition does not take into account the frequency of occurrence of collocations to consider them as such. The annotation of learner texts involved the classification of collocation errors, according to a typology devised by Alonso Ramos et al. (2010b). The correctness of collocations produced by learners was established based on native speaker intuition as well as corpus data (if a collocation was deemed incorrect by the annotators, but presented at least five occurrences in the *Corpus de Referencia del Español Actual* [http://corpus.rae.es/creanet.html], it was considered correct).

The full annotation scheme makes possible to retrieve information on collocation errors, as well as to group collocations according to their syntactic pattern and the meaning of the collocate (e.g. collocations made up of a delexical verb plus its object, collocations that consist of a noun modified by an intensifier adjective, etc.). These groupings allow for a study that would be impossible taking into account particular collocations. For instance, it seems futile to make generalizations based on occurrences of a specific collocation such as *dar un paseo* ('to take a walk'), since we will not be able to find many instances of this collocation, given the limited overall number of occurrences in the corpus. However, if we focus on more abstract entities, such as the group of collocations formed by delexical verbs combined with predicate nouns, we can obtain a considerable amount of data.

Even though some findings based on the data resulting from this annotation process (types of errors found, rate of lexical vs. grammatical errors, influence of the L1, etc.) have been already presented elsewhere (Alonso Ramos et al. 2010a, Vincze et al. 2011, Wanner et al. 2013), there are still several aspects of learners' collocation use that have not been dealt with, such as:
a) What semantic types and syntactic patterns of collocations are most frequently used by learners and native speakers? There is any difference between these two groups?
b) Do different collocational types present different error rates (i.e. do learners make more errors when producing, for instance, intensifying collocates than evaluative collocates)? How are these error rates distributed?
c) Does the frequency in a general corpus (hypothetically representing frequency in learners' input) have any relationship with the error rate displayed by a given collocation type?

In order to answer the questions in (a), collocations found in the learner corpus have been compared to those found in a native corpus (also a section of CEDEL2) to establish whether there are differences regarding the frequency of use of different collocation types by native and non-native speakers. The existence of differences shows that learners overuse certain groups of collocations, but there is evidence suggesting a lack of lexical richness on their part. As for question (b), different error rates in the case of different collocation types can indicate that certain groups of collocations are more difficult than others for learners of Spanish. Finally, the answer to question (c) involves the comparison between the correct and incorrect collocations found in the learner corpus in order to determine whether the collocations attempted when incorrect collocations are produced are less frequent than the collocations produced correctly by learners in the input received. This implies to obtain frequency information from a Spanish reference corpus.

The three factors analysed can be useful in predicting the difficulty of collocations and, therefore, in considering their treatment in second language curricula.

References

Alonso Ramos, Margarita, Leo Wanner, Nancy Vázquez Veiga, Orsolya Vincze, Estela Mosqueira Suárez and Sabela Prieto González. 2010a. "Tagging Collocations for Learners." In *Elexicography in the 21st Century: New Challenges, New Applications. Proceedings of eLex2009*, edited by Sylviane Granger and Magali Paquot (375–80). Louvain-la Neuve, Presses Universitaires de Louvain: Cahiers du cental 7.

Alonso Ramos, Margarita, Leo Wanner, Orsolya Vincze, Gerard Casamayor, Nancy Vázquez Veiga, Estela Mosqueira Suárez, and Sabela Prieto González. 2010b. "Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora." In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), 19-21 May 2010, Valletta, Malta*, edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias (3209–14). ELRA.

Hausmann, Franz Josef. 1989. "Le dictionnaire de collocations. In *Wörterbücher, Dictionaries, Dictionnaires*, Vol. 1., edited by Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand & Ladislav Zgusta (1010–19). Berlin: de Gruyter.

Lozano, Cristobal, and Amaya Mendikoetxea. 2013. "Learner Corpora and Second Language Acquisition: The Design and Collection of CEDEL." In *Automatic Treatment and Analysis of Learner Corpus Data*, edited by A. Díaz-Negrillo, N. Ballier, and P. Thompson (65–100). Amsterdam: John Benjamins.

Mel'čuk, Igor. 1996. "Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon." In *Lexical Functions in Lexicography and Natural Language Processing*, edited by Leo Wanner (37–102). Amsterdam/Philadelphia: John Benjamins.

Vincze, Orsolya, Margarita Alonso Ramos, Estela Mosqueira Suárez, and Sabela Prieto González (2011). "Exploiting a Learner Corpus for the Development of a CALL Environment for Learning Spanish Collocations." In *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011*, edited by Iztok Kosem and Karmen Kosem, (280–85). Ljubljana: Trojina, Institute for Applied Slovene Studies.

Wanner, Leo, Serge Verlinde and Margarita Alonso-Ramos. 2013. "Writing assistants and automatic lexical error correction: Word combinatorics". In *Proceedings of eLex 2013: electronic lexicography in the 21st century: thinking outside the paper*, edited by Itzok Kosem, Jelena Kallas, Polona Gantar, Simon Krek, Margit Langemets, Maria Tuulik (472-487). Tallin, Estonia.