Identification of Cross-disciplinary Spanish Academic Collocations for a Lexical Tool

Margarita Alonso Ramos, Marcos García Salido,

Marcos Garcia, Eleonora Guzzi

Universidade da Coruña, CITIC, Grupo LyS,
Dpto. de Letras, Fac. de Filoloxía. 15071, A Coruña
E-mail: {margarita.alonso, marcos.garcias, marcos.garcia.gonzalez}@udc.ga,
eleonora.guzzi@udc.es

We aim to build a lexical tool that helps novice writers in their academic writing in Spanish (Alonso-Ramos et al., 2017). Although most academic texts at Spanish universities are written in Spanish and Spanish is the mother tongue of the majority of students, the latter does not guarantee a good writing performance in academic discourse. Academic writing has to be learnt, since there is no native speaker of this genre. In fact, the academic writing of university students often shows certain deficiencies, many of which come from a poor knowledge of collocations. The proposed lexical tool offers suggestions of Spanish cross-disciplinary collocations, in order to help university students by improving the quality of their academic lexicon (for further details, see also García-Salido et al., 2018).

We focus on the proper method to identify cross-disciplinary collocations in a Spanish academic corpus consisting of research articles. Even though there are important lexical differences in different domains (Hyland & Tse, 2007), our project follows the approach according to which specialized texts contain, besides general lexicon (Drouin, 2007; Jacques & Tutin 2018: 1) domain-specific lexicon (or terminology) and 2) cross-disciplinary lexicon (or academic lexicon), which is in line with several works on academic English (Coxhead, 2000, Ackermann & Chen 2013; Gardner & Davies, 2014, Frankenberg-Garcia et al., 2018). However, the distinction between both kinds of lexicon is not clear-cut, especially when we deal with collocations. It is not enough to verify that the two elements of collocations are sufficiently represented in different domains of the academic corpus separately, but also the collocation as a whole. For instance, the noun *actividad* 'activity' and the verb *presentar* 'to present' have been selected for their specificity in the academic corpus, but the collocation *presentar actividad* is only specific to the domain of Natural Sciences.

We will describe the process of extraction of collocations from our academic corpus and the process of manual filtering that we employed until now. Firstly, we extracted a list of academic word candidates based on their specificity and on dispersion across all the domains. Secondly, we parsed our academic corpus to build a list of word combinations

using syntactic dependencies. From the 418 collocation candidates corresponding to 38 bases, we manually filtered those which were proper collocations. Out of these, only 113 collocations (from 25 bases) that were considered cross-disciplinary have been selected. The other 305 collocation candidates have been discarded mainly because they were considered free phrases or terminological. In order to improve the efficiency of this manual filtering, we compiled a bigger domain-specific corpus using WebBootCat (Baroni et al., 2006) with four main domains and 12 subdomains. After some experiments, we applied the Inverse Document Frequency model, a dispersion measure, to verify if a collocation is significantly more frequent in a given subdomain. If so, it will not be considered cross-disciplinary. We will present the results of these experiments, as well as the current state of the collocational tool.

Keywords: cross-disciplinary; academic vocabulary; collocations

References

Ackermann, K., & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), pp. 235-247.

Alonso-Ramos, M., García Salido, M. & Garcia, M. (2017). Exploiting a corpus to compile a lexical resource for academic writing: Spanish lexical combinations. In I. Kosem et al. (eds) *Proceedings of 2017 eLex Conference*, Leiden. pp. 571-586.

Baroni, M., Kilgarriff, A., Pomikálek, J. & Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In *Proceedings of EAMT 2006.* Oslo, pp. 247-252.

Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, *34*(2), pp. 213-238.

Drouin, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, *12*(2), pp. 45-64.

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P. & Sharma, N. (2018). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, *31*(1), pp. 1-17.

García-Salido, M., Villayandre, M. & Alonso-Ramos, M. (2018). A Lexical Tool for Academic Writing in Spanish based on Expert and Novice Corpora. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), pp. 260-265.

Gardner, D, & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, *35*(3), pp. 305-327.

Hyland, K. & Tse's, P. (2007). Is there an "academic vocabulary"?. *TESOL quarterly*, *41*(2), pp. 235-253.

Jacques, M. P. & Tutin, A. (2018). *Lexique transversal et formules discursives des sciences humaines*. London: ISTE Editions.