Research results and outcomes of the project “A Phraseographical Methodology and Model for an Online Corpus-Based Multilingual Collocations Dictionary Platform”

This presentation aims to show some results and outcomes achieved within the scope of a project funded by The São Paulo Research Foundation (Process 2020/01783-2). It had the purpose of developing a phraseographical methodology and model for an online corpus-based Multilingual Collocations Dictionary Platform (MULTPLATCOL), in English, Portuguese, French, Spanish, and Chinese. The follow-up proposal will also include Italian, German, European Portuguese and Arabic. MULTPLATCOL is aimed to be customised for different target audiences according to their needs: language learners, pre- and in-service teachers, translators, material developers and researchers or lexicographers (Bothma; Tarp: 2012, Fuertes-Olivera; Tarp: 2014; Tarp: 2015). In this talk, we will present and discuss some data results and analysis in English, Portuguese, Spanish and Chinese, related to the project methodology, and the development of an in-house Collocations Dictionary Writing System (COLDWS).

The methodology developed for the MULTPLATCOL relies on the combination of automatic methods to extract candidate collocations (Garcia et al. 2019a). The automatic approaches take advantage of NLP tools to annotate large corpora with lemmas, PoS-tags and dependency relations in the five languages. Using these data, we applied statistical measures (Evert et al. 2017; Garcia et al. 2019b) and distributional semantics strategies to select the collocation candidates (Garcia et al. 2019c) and retrieve corpus-based examples (Kilgarriff et al. 2008), and we followed Garcia et al. (2019c) to carry out an automatic translation of the collocations (Undisclosed references, 2021). All automatically extracted data have been carefully post-edited by the lexicographers involved in this investigation. The results regarding, for example, the number of extracted collocation candidates in the first phase is significantly high (a total of 309,838 collocation candidates) and demands a lot of effort from the whole team to carefully review them.

Regarding the aforementioned methodological aspects, they have all proved to be efficient for all languages, except for Chinese. The treatment of the Chinese has created several difficulties due to the morphological and grammatical particularities of this language. For example, in the automated process of most languages, a simple question on how to delimit the extension of a word can raise unsuspected problems in Chinese, where the indispensable task of segmentation is not completely trivial (Undisclosed references, in press).

With respect to COLSDWS, we have also developed this software aimed at specifically compiling and producing collocation dictionaries. This way, all automatically extracted data was automatically inserted into the COLDWS, post-edited by the lexicographers, and will be afterwards exported to an end-user platform. The COLSDWS, which will also be presented in this paper, is one of the practical and satisfactory outcomes of this project, being fully operational and particularly effective. An end-user platform model, which will be functionally integrated with the COLSDWS, was also designed and is in
process of adjustments. It still requires some improvements to adjust it to our goal of building a platform that generates a more ambitious, customised and interactive lexicographic work, more suitable to the specificities and the idiosyncrasies of its users and their lexicographic needs.

**Keywords**: collocations; collocations dictionary; multilingual platform; Dictionary Writing System

**References**


Undisclosed references. (2021)
Undisclosed references. (in press)
