Adriane Orenha Ottaiano/Maria Eugênia Olímpio de Oliveira Silva/Carlos Roberto Valêncio/João Pedro Quarado

DEVELOPING A COLLOCATION DICTIONARY WRITING SYSTEM (COLDWS) FOR AN ONLINE MULTILINGUAL COLLOCATIONS DICTIONARY PLATFORM (PLATCOL)

Keywords Dictionary writing system; collocations dictionary; multilingual dictionary; collocations

This ongoing research project, funded by the São Paulo Research Foundation (FAPESP – 2020/01783-2), has the purpose of developing a phraseographical methodology and model for an Online Corpus-Based Multilingual Collocations Dictionary Platform (PLATCOL), in five different languages so far: English and Portuguese, French, Spanish, and Chinese. It is aimed to be customized for different target audiences according to their needs, such as language learners, pre- and in-service teachers, translators, material developers and researchers or lexicographers. To achieve this goal, we follow the theoretical assumptions of the function theory of lexicography (Bothma/Tarp 2012; Fuertes-Olivera/Tarp 2014; Tarp 2015). Hence, both the procedures chosen for the selection, organization and presentation of lexicographic data, as well as the determination of the content, form and access routes are adapted and subordinated to the users' preferences.

The methodology being developed for the PLATCOL 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 apply statistical measures (Evert et al. 2017; Garcia/García-Salido/Alonso-Ramos 2019b) and distributional semantics strategies to select the collocation candidates (Garcia/García-Salido/Alonso-Ramos 2019c) and to retrieve corpus-based examples (Kilgarriff et al. 2008). Having the lexicographers selected the suitable collocations, we follow Garcia/García-Salido/Alonso-Ramos (2019c) to carry out an automatic translation of the collocations. All automatically extracted data are being carefully post-edited by the lexicographers involved in this investigation (Orenha-Ottaiano et al. 2021).

In order to better enable or enhance the post-editing of all the automatically retrieved data, we developed an in-house Collocations Dictionary Writing System (COLDWS), a software aimed at specifically compiling and producing collocation dictionaries, so that all automatically extracted data can be automatically inserted into this COLDWS, post-edited by the lexicographers, as well as be afterwards exported to an end-user platform.

In this paper, we will focus on the theoretical as well as methodological aspects regarding the development of the COLDWS, duly created to fulfill the specific needs of our collocations dictionaries. We have knowledge of the dictionary writing system Lexonomy, a web-based dictionary writing system (Měchura 2017), and TshwaneLex (de Schryver 2007), two good quality DWSs. However, as previously said, we needed to rely on a DWS that would meet the specificities of a multilingual collocations dictionary and also deal with the specific data output generated for our project.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The COLDWS exclusively focuses on the management of entries, collocations and all the other dictionary data related to them, automatically retrieved and automatically inserted into this software. With this software, it is possible to register and edit all dictionary information, such as languages, corpora data, morphosyntactic structures, taxonomy of the collocations, translations, statistical measures etc. There is also a functionality with which reviewers can post-edit entries, collocations and translations. With respect to the validation process, the software will allow lexicographers to choose from three phases (traffic lights phases), indicating to users the status of the entries or collocations. In the first phase, data automatically inserted into the COLDWS (not revised yet) will be displayed with a red icon, even to users, when exported to the end-user platform. Phase 2 represents data revised by one member of the team (reviewer 1), but which may still need a second evaluation and/or some adjustments – an orange icon will be shown. In phase 3, data is checked by a second reviewer (reviewer 2) and now considered to be suitable – a green green icon will be displayed.

Besides that, the COLDWS will also be of valuable help when it comes to optimizing translation of entries or collocations. For example, once translation pairs between collocations are identified and registered in the system, making up a multilingual database, it becomes possible to identify and automatically suggest new translations among other languages. This process occurs through an inference-based algorithm, built from an inference hypothesis related to the composition of multiple translation dictionaries: if word or collocation A translates into word or collocation B which in turn translates into word or collocation C, what is the probability that C is a translation of A? Studies developed under this hypothesis (e. g. Mausam et al. 2010) presented significant results in relation to the analysis via inference of translation pairs between different languages. In this process, the algorithm performs the analysis of previously registered translations, identifies other translation pairs via inference, and shows lexicographers the possibilities of translations, who must analyze the reliability and quality of the translation found. To our knowledge, there is not any DWS functionality compared to this one and that may be a successful innovation with respect to DWS functionalities, enhancing the development of collocations dictionaries.

In what concerns computational development, the COLDWS was built using languages from current and widespread programming such as Java (with Model View Controller [MVC] architecture), HTML and Javascript (jQuery library). For the storage, consultation and deletion of entries, the relational data model with PostgreeSQL software was used in conjunction with SQL language. In addition, concepts of User Experience – UX were applied to provide a good experience for the lexicographers.

The DWS is still under evaluation and if the results are positive, we will also allow free access to and use of the software upon request.

References

Bothma, T. J. D./Tarp, S. (2012): Lexicography and the relevance criterion. In: Lexikos 22, pp. 86–108.

de Schryver, G.-M./De Pauw, G. (2007): Dictionary writing system (DWS) + corpus query package (CQP): the case of Tshwane Lex. In: Lexikos 17, pp. 226–246.

Evert, S./Uhrig, P./Bartsch, S./Proisl, T. (2017): E-VIEW-affilation – a large-scale evaluation study of association measures for collocation identification. In: Kosem, I./Tiberius, C./Jakubíček, M./Kallas, J./Krek, S./Baisa, V. (eds.): Electronic Lexicography in the 21st Century: Lexicography

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

from Scratch. Proceedings of eLex 2017. Leiden, the Netherlands, 19–21 September 2017. Brno, pp. 531–549. https://elex.link/elex2017/proceedings/eLex_2017_Proceedings.pdf (last access: 12-03-2020).

Fuertes Olivera, P. A./Tarp, S. (2014): Theory and practice of specialised dictionaries. Lexicography versus terminography, Berlin/Boston.

Garcia, M./García-Salido, M./Alonso-Ramos, M. (2019a): Towards the automatic construction of a multilingual dictionary of collocations using distributional semantics. In?: Kosem, I./Kuhn,T. Z./ Correia, M./Ferreira, J. P./Jansen, M./Pereira, I./Kallas, J./Jakubíček, M./Krek, S./Tiberius, C. (eds.): Electronic Lexicography in the 21st Century. Proceedings of eLex 2019. Sintra, Portugal, 1–3 October 2019. Brno, pp. 747–762. https://elex.link/elex2019/wp-content/uploads/2019/09/ eLex_2019_42.pdf (last access: 10-03-2020).

Garcia, M./García-Salido, M./Alonso-Ramos, M. (2019b): A comparison of statistical association measures for identifying dependency-based collocations in various languages. In: Savary, A./Parra Escartín, C./Bond, F./Mitrović, J./Mititelu, V. B. (eds.): Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019). Florence, Italy, August 2, 2019, Association for Computational Linguistics. Stroudsburg, pp. 49–59. https://www.aclweb.org/anthology/W19-5107. pdf (last access: 09-03-2020).

Garcia, M./García-Salido, M./Alonso-Ramos, M. (2019c): Weighted compositional vectors for translating collocations using monolingual corpora. In: Corpas Pastor, G./Mitkov, R. (eds.) Computational and corpus-based phraseology. Cham, pp. 113–128.

Kilgarriff, A./Husák, M./McAdam, K./Rundell, M./Rychly, P. (2008): GDEX: automatically finding good dictionary examples in a corpus. In: Bernal, E./DeCesaris, J. (eds.): Proceedings of the 13th EURALEX International Congress. Barcelona, 15–19 July 2008. Barcelona, pp. 425–432.

Mausam, S./Etzioni, S./Weld, O./Reiter, D. S./Skinner, K./Sammer, M./Vessier, S. (2010): Panlingual lexical translation via probabilistic inference. In: Artificial Intelligence 174 (9–10), pp. 619–637.

Měchura, M. B. (2017): Introducing lexonomy: an open-source dictionary writing and publishing system. In: Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of eLex 2017. Leiden, the Netherlands, 19–21 September 2017. Brno, pp. 662–679.

Orenha-Ottaiano, A./Garcia-Gonzalez, M./Olímpio, M. E./L'Homme, M./Alonso Ramos, M./Valencio, C. R./Tenorio, W. (2021): Corpus-based methodology for an online multilingual collocations dictionary: first Steps. In: Kosem, I./Cukr, M./Jakubíček, M./Kallas, J./Krek, S./Tiberius, C. (eds.): Electronic Lexicography in the 21st Century. Proceedings of the eLex 2021 conference. 5–7 July 2021, virtual. Brno, pp. 1–28. https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_01_ pp1-28.pdf (last access: 04-08-2021).

Tarp, S. (2015): La teoría funcional en pocas palabras. In: Estudios de Lexicografía 4, pp. 31–42.

Contact information

Adriane Orenha-Ottaiano
São Paulo State University (UNESP)
adriane.ottaiano@unesp.br

Maria Eugênia Olímpio de Oliveira Silva
University of Alcalá
eugenia.olimpio@uah.es

Carlos Roberto Valêncio
São Paulo State University (UNESP)
carlos.valencio@unesp.br

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/ Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

João Pedro Quadrado
São Paulo State University (UNESP)
jp.quadrado@unesp.br

Acknowledgments

We gratefully acknowledge the financial support provided by The São Paulo Research Foundation (FAPESP), Process ner 2020/01783-2.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.