Development of a methodology and enhancements of lexicographical resources for an online Platform of Academic Collocations Dictionaries in Portuguese and English

Adriane Orenha-Ottaiano [1], Tanara Zingano Kuhn [2], Carlos R. Valêncio [1], João P. Quadrado [1], Stella E. O. Tagnin [3], Arnaldo Candido Júnior [1]

[1] São Paulo State University (UNESP), Brazil
[2] CELGA-ILTEC, Universidade de Coimbra, Portugal
[3] University of São Paulo (USP), Brazil

adriane.ottaiano@unesp.br; tanarazingano@outlook.com; carlos.valencio@unesp.br; jp.quadrado@unesp.br; seotagni@usp.br; arnaldo.candido@unesp.br

Motivation

In view of the growing demand for publications of academic texts and for the dissemination of studies in national and international congresses, the development of specific lexicographic tools that can assist with academic writing is fundamental. Some resources are already available, such as dictionaries of academic language (e.g., Oxford Learner's Dictionary of Academic English; The Louvain EAP Dictionary (https://leaddico.uclouvain.be/) and writing assistant tools that have a lexicographic component (e.g., ColloCaid, Frankenberg-Garcia et al., 2019; HARTAes-vas, Alonso-Ramos & Zabala, 2022; Gracia-Salido et al., 2018). As can be seen, some languages are better equipped with this type of resources than others. However, dictionaries whose unique focus is on academic collocations are, to the best of our knowledge, still unheard of in any language.

Objectives

The objective of our research project is thus to contribute to filling this gap by creating *Online Dictionaries of Academic Collocations*.

Initially, English and Brazilian Portuguese will be the languages covered, but we intend to include more languages in next phases of the project.

During this three-year of a publicly funded research project (CNPq, Process nr. 409178/2021-7), our main objectives are:

- to develop a methodology for the creation of corpus-driven academic collocation dictionaries
- to improve an existing dictionary writing system and an end-user interface (*PLATCOL*, Orenha Ottaiano, 2020; Orenha Ottaiano et al., 2021a; Orenha Ottaiano & Silva 2021b) for the purposes of this project.

Academic Collocations

Even though academic collocations have received considerable attention in the past years, they still lack a more suitable and explicit definition, due to its complexity.

Reference	Definition/methodology
Paquot, M. (2010). A data-driven approach to the selection of academic vocabulary. In Paquot, M. (2010). *Academic Vocabulary in Learner Writing: from extraction to analysis*, p. 29-63.	- 'potential academic words' - used to refer to words that are reasonably frequent in a wide range of academic texts but relatively uncommon in other kinds of texts and which, as such, might be used to refer to those activities that characterize academic work, organize scientific discourse and build the rhetoric of academic texts, and so be granted the status of academic vocabulary.
Ackerman, K. & Chen, H. Y. (2013). Developing the Academic Collocation List (ACL) - A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12(4), 235-247.	- list of node words: the words that occurred at least five times per million words and in at least five different texts in the written curricular component of the *Pearson International Corpus of Academic English* (PICAE) comprising over 25 million words. Each of the four fields of study contains materials from seven academic disciplines to ensure that the corpus is representative of the academic register. - Words from the General Service List (West,1953) were removed from the node word list but could appear as pre- or post-collocate.
Durrant, P. (2009). Investigating the viability of a collocation list for students of English for Academic Purposes. *English for Specific Purposes*, 28(3): 157-169.	- A number of researchers are currently attempting to create listings of important collocations for students of EAP. However, so far these attempts have (1) failed to include positionally-variable collocations, and (2) not taken sufficient account of variation across disciplines.

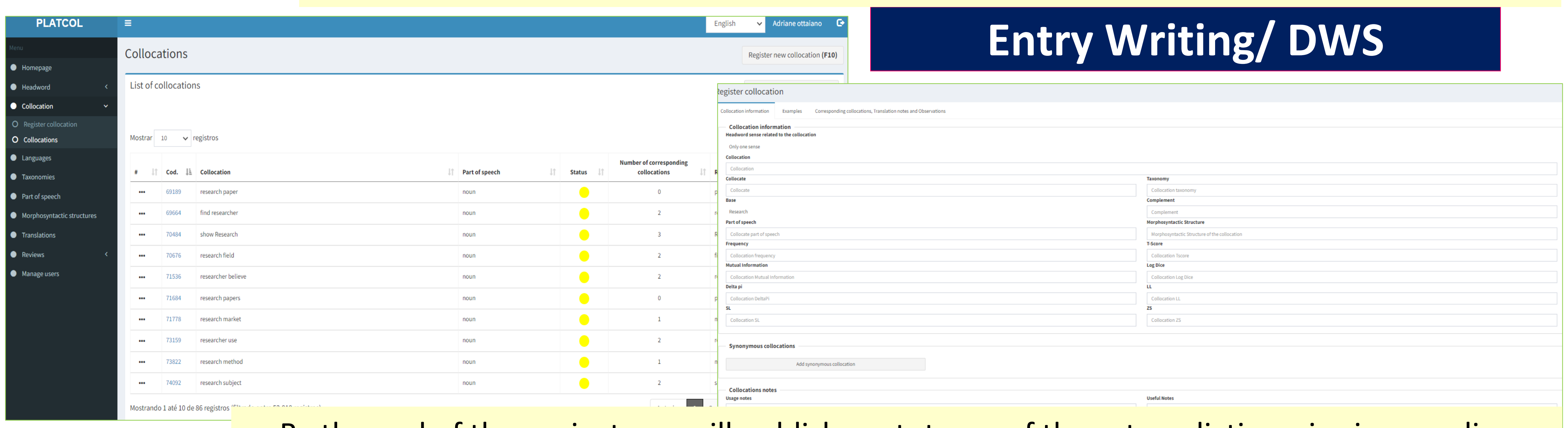
Our understanding – General Collocations X Academic Collocations

General Collocations	Academic Collocations
- Pazos-Bretaña, Orenha-Ottaiano and Xiong (in press): *Under a statistically oriented approach, we view collocations as frequent word combinations whose co-occurrence within a certain distance of each other is statistically higher than expected in comparison to any other words randomly combined in a specific language (Barfield; Gyllstad, 2009; Nesselhauf, 2005; Sinclair 1966, 1991; etc.).* ➢ *However, as Teubert (2004: 188) mentioned being statistically significant is not enough to identify a combination of words as a collocation: 'They also have to be semantically relevant. They have to have a meaning of their own, a meaning that isn't obvious from the meaning of the parts they are composed of'.* - *For this reason, it is important to describe collocations under a phraseological approach, and so we define collocations as pervasive, recurrent, and conventionalized combinations consisting of a base and a collocate (Haussmann 1979, 1989), which are lexically and/or syntactically fixed to a certain degree.* *They can be said to be partially compositional because its base maintains its meaning, however, the collocate may take on a special meaning only when combined with the base (Alonso-Ramos, 1994; Corpas 1996; Haussmann, 1989; Heylen; Maxwell, 1994; Orenha-Ottaiano, 2020; Pamies, 2019; Penadés Martínez, 2017; Torner & Bernal, 2017)*	- Under a statistically oriented approach, we view academic collocations as frequent word combinations in academic texts, whose co-occurrence within a certain distance of each other is statistically higher than expected in comparison to any other words randomly combined in a specific language and in a specific field of knowledge. - Under a phraseological approach, academic collocations are combinations of words that are recurrent and conventionalized in academic texts, which may have taken on a different or new meaning from the ones used in non-academic language, being varied across disciplines. Like the general collocations, they also consist of a base and a collocate which are combined to structure these types of texts in the different fields of knowledge. They can also be partially compositional - *the collocate may take on a special meaning only when combined with the base* - and they can be identified on a continuum, ranging from free to more restricted combinations, according to the field of knowledge they may be in.

Methodology

- Brazilian subcorpus of the *Corpus of Portuguese from Academic Journals* (Kuhn & Ferreira, 2020) - 6 areas of knowledge (20 million words)
- The Sketch Engine
- Academic Collocations identification: corpus-based and corpus-driven
  - ➢ The methodology and decisions for the extraction will be defined as we move forward with data analysis. Decision 1 – words considered core general vocabulary should not be excluded from the list of nodes

Entry Writing/ DWS



Collocations

List of collocations

Mostrar 10 registros

#	Cod.	Collocation	Part of speech	Status	Number of corresponding collocations
•••	69189	research paper	noun	●	0
•••	69664	find researcher	noun	●	2
•••	70484	show Research	noun	●	3
•••	70676	research field	noun	●	2
•••	71536	researcher believe	noun	●	2
•••	71684	research papers	noun	●	0
•••	71778	research market	noun	●	1
•••	73159	researcher use	noun	●	2
•••	73822	research method	noun	●	1
•••	74092	research subject	noun	●	2

Mostrando 1 até 10 de 86 registros

Register collocation

Collocation information | Examples | Corresponding collocations, Translation notes and Observations

Collocation information

Headword sense related to the collocation: Only one sense
Collocation: Collocation
Collocate: Collocate
Taxonomy: Collocation taxonomy
Base: Research
Complement: Complement
Part of speech: Collocate part of speech
Morphosyntactic Structure: Morphosyntactic Structure of the collocation
Frequency: Collocation frequency
T-Score: Collocation Tscore
Mutual Information: Collocation Mutual Information
Log Dice: Collocation Log Dice
Delta pi: Collocation DeltaPi
LL: Collocation LL
SL: Collocation SL
ZS: Collocation ZS

Synonymous collocations

Add synonymous collocation

Collocations notes

Usage notes
Useful Notes

Future Work

By the end of the project, we will publish prototypes of these two dictionaries in an online platform. With a tested methodology of dictionary-making and fully functional adaptations to the dictionary writing system and end-user interface, we will be able to move to a new phase in which not only will we work on turning the prototypes into fully fledged dictionaries, but we will also be able to include additional languages.

Team and Collaborators

- Dr. Adriane Orenha-Ottaiano (Coordinator) – São Paulo State University (UNESP), Brazil
- Dr. Tanara Zingano Kuhn (Vice-Coordinator) – University of Coimbra, Portugal
- Dr. Carlos Roberto Valêncio – UNESP, Brazil
- Dr. Cristiane Krause Kilian – Instituto Sup. de Educação Ivoti, Brazil
- Dr. Arnaldo Candido Júnior – UNESP, Brazil
- Dr. Ivan Rizzo Guilherme – UNESP, Brazil
- Dr. Stella E. Ortweiler Tagnin – University of São Paulo, Brazil
- Dr. Nils Reiter – University of Cologne, Germany
- Francisco Mondaca – Research Associate, University of Cologne, Germany
- Dr. Iztok Kosem (Consultant) – University of Ljubljana, Slovenia/Jozef Stefan Institute
- Dr. Špela Arhar Holdt (Consultant) – University of Ljubljana, Slovenia

References

Alonso-Ramos, M. & Zabala, I. (2022). HARTAes-vas: Lexical Combinations for an Academic Writing Aid Tool in Spanish and Basque. SEPLN (Projects and Demonstrations), pp. 22-25.

Frankenberg-Garcia, A. Lew, R., Roberts, J., Rees, G. & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, 32/2, pp. 23-39.

García-Salido, M., M. Garcia, M. Villayandre & M. Alonso-Ramos. 2018. A Lexical Tool for Academic Writing in Spanish based on Expert and Novice Corpora. In N. Calzolari et al. (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), pp. 260–265.

Granger, S. & Paquot, M. *The Louvain EAP Dictionary*. Available at: https://leaddico.uclouvain.be/. (last access: 27-01-2023).

Kuhn, Tanara Zingano; Ferreira, José Pedro. (2020). O Corpus de Português Escrito em Periódicos - CoPEP. *DELTA: Documentação e Estudos em Linguística Teórica e Aplicada*, 36(2).

Orenha-Ottaiano, A.; Garcia, M.; Olímpio De Oliveira, M. Eugênia; L'Homme, M-C; Alonso Ramos, M.; Valêncio, C. R. & Tenório, W. (2021a). Corpus-based methodology for an Online Multilingual Collocations Dictionary: First Steps. In: Kosem, I.; Michal C.; Miloš J.; Jelena K.; S. Krek & C. Tiberius (eds.). *Proceedings of eLex 2021*, pp. 1-28.

Orenha-Ottaiano, A.; Silva, M. E. O. O. (2021b). A Corpus-based Platform of Multilingual Collocations Dictionaries (PLATCOL): some lexicographical aspects aiming at pre- and in-service teachers. *Proceedings of the International Conference Corpus Linguistics 2021*. St. Petersburg, Russia, pp. 122-132.

Orenha-Ottaiano, A. (2020). A phraseographical methodology and model for an online corpus-based multilingual collocations dictionary platform. Research Grant (2020-2022) awarded by *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP Process ner. nº 2020/01783-2).

Oxford Learner's Dictionary of Academic English. (2014). Oxford: Oxford University Press.

Pazos Bretaña, J. M.; Orenha-Ottaiano, A. & Xiong, Z. (in press). PLATCOL, Plataforma Multilingüe de Diccionarios de Colocaciones: el caso del chino. Estudios de Traducción.

Acknowledgements: The authors gratefully acknowledge the financial support provided *by The Brazilian National Council for Scientific and Technological Development* (*CNPq*), Process nr. 409178/2021-7, Brazil, and the Portuguese National Funding Agency, FCT - Foundation for Science and Technology, I.P. (grant number UIDP/04887/2020).