

## DISSERTATION ABSTRACT

# A Design Proposal of an Online Corpus-Driven Dictionary of Portuguese for University Students

Tanara Zingano Kuhn

University of Lisbon, PT

[tanarazingano@outlook.com](mailto:tanarazingano@outlook.com)

---

The objective of this PhD project was to propose the design of an online corpus-driven dictionary of Portuguese for university students (DOPU), aimed at both speakers of Portuguese as a mother tongue and as an additional language and covering Brazilian and European Portuguese varieties. For that, the highly innovative semi-automated approach to dictionary-making (Gantar, Kosem and Krek 2016) was adopted, which involves automatic extraction of data from the corpus and import into dictionary writing system. As a method that had never been applied for lexicographical projects of the Portuguese language, it was necessary to experiment the approach for the first time. Thus, all the required pre-requisites were newly developed, namely, a corpus of academic texts, sketch grammar, GDEX configuration, and a specially-tailored procedure for automatic extraction of data. The experiment indicated that not only can this approach be successfully used as a means to provide lexical content for the design of DOPU, but it can also be beneficial to other lexicographical projects of Portuguese.

---

**Keywords:** academic Portuguese; automated lexicography; corpus; dictionary; tool development

---

University students are expected to read and write academic texts as part of typical literacy practices in higher education settings. One of the conditions for students to engage in these routine activities involves “learning to use language in new ways” (Hyland 2009: viii–ix). Among many pedagogical resources that can help students reach academic language proficiency, dictionaries are undoubtedly one of the most fundamental. While for many languages, dictionaries of academic language have been created, Portuguese still lacks such an important pedagogical resource.

In order to support the mastery of academic Portuguese, this PhD research proposed a design of an online corpus-driven dictionary of Portuguese for university students (DOPU) attending Portuguese-medium institutions, speakers of Brazilian Portuguese (BP) and European Portuguese (EP), both as a mother tongue and as an additional language. At this point, special attention should be drawn to the fact that the term *design* was used as defined in Hartmann and James’ *Dictionary of Lexicography* (1998): “**design**. The overall principles that govern the production of efficient REFERENCE WORKS, taking into account not only features of content (INFORMATION CATEGORIES) and presentation (ARRANGEMENT), but also the reference needs and skills of the USER.”

One of the key parts of the process of creating principles that regulate dictionary making concerns defining the methodology. In the case of this proposal, it was decided to experiment, for the first time for the Portuguese language, the semi-automated approach to dictionary making, as originally put forward by Rundell and Kilgarriff (2011) and first implemented into lexicographic practice by Gantar, Kosem, and Krek (2016) in a project for making a dictionary of Slovene. In this highly innovative approach, lexical data are automatically extracted from the corpus according to predetermined criteria and

transferred to the dictionary writing system (DWS), where lexicographers then analyse, validate and edit the data to shape them into the final database entry. In this thesis, this procedure was performed on the Sketch Engine corpus tool (Kilgarriff et al. 2004) and the dictionary writing system used was iLex (Erlandsen 2010).

Given that such a method had never been applied to Portuguese, the initial step involved verification of the availability of the resources that were required for the procedure, namely, a sketch grammar, GDEX configuration, a corpus of academic texts, and procedure (including an API script) for automatic extraction of data and import into the DWS. Due to unsuitability of the first two resources and inexistence of the other two, it was decided that those pre-requisites had to be especially developed.

Firstly, a new corpus was compiled, the *Corpus de Português Escrito em Periódicos* – CoPEP (‘Corpus of Portuguese from Academic Journals’). This corpus contains around 10,000 texts totalling over 40 million tokens extracted from academic journals published in the Brazilian and Portuguese national collections of SciELO (Scientific Electronic Library Online), distributed among three schools of knowledge, and further divided into six great areas, following the CAPES classification of areas of knowledge (Brazil).

Secondly, a new sketch grammar for Portuguese was devised. A sketch grammar is a file with grammatical relations and processing directives for the Sketch Engine system to compute different types of relations through statistical calculations. The data obtained from these computations then form the basis of the word sketch feature in the Sketch Engine, which is the heart of the process of automatic extraction of data from the corpus.

Thirdly, GDEX (Good Dictionary Examples, Kilgarriff et al. 2008) configurations for Portuguese were further developed, considering the corpus in question, i.e., CoPEP, and the purpose of the examples, i.e., to be used for writing entries for DOPU. GDEX is a function in the Sketch Engine tool that, based on pre-defined criteria, identifies example sentences in the corpus, placing the best ones at the top of the list of concordance lines in order to facilitate the lexicographer’s process of example selection.

Finally, the extraction procedure developed by Gantar, Kosem, and Krek (2016) was adapted, with two additions that were specifically developed for this thesis. The first one was the inclusion of additional information provided by the clustering and longest-commonest match (Kilgarriff et al. 2015) functions in the Sketch Engine. This information was added to the data after the extraction, at a post-processing stage. The main aim was to assist lexicographers in grouping collocates and in identifying multi-word expressions, as well as to facilitate the detection of incorrect information.

While the first addition was language non-specific, i.e., it can be used in automatic extraction for other languages, the second was specific to Portuguese: assignment of Portuguese variety labels not only to headwords, but also to collocations, and if relevant, to grammatical relations. This addition stems from the fact that DOPU aims to equally represent the two varieties of Portuguese. Considering that the original procedure was meant for Slovene, which is a monovarietal language, the pluricentric nature of Portuguese posed completely new challenges for data extraction.

Evaluation of the adoption of the semi-automated approach in the context of the DOPU design indicated that although further development of these brand-new resources and tools, as well as the procedure itself, would greatly contribute to increasing the quality of DOPU’s lexical content, the extracted data can already be used as a basis for entry writing. This means that the design can be already implemented, thus making DOPU immediately available, initially as a work-in-progress resource, with entries containing automatically extracted collocations and examples to which students can already have access. The positive results of the experiment also suggest that this approach should be highly beneficial to other lexicographic projects of Portuguese as well.

## Acknowledgements

I am grateful to the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (Capes-Brazil) for the PhD scholarship (process number 0973/13-0), to CELGA-ILTEC (University of Coimbra) for funding essential elements of my PhD research, to the European Cooperation for Science and Technology (COST) through the European Network of e-Lexicography (ENeL) Action for a grant (COST-STSM-IS1305-210216-071459) for a Short-Term Scientific Mission at the University of Ljubljana, and to the University of Ljubljana for granting me licence to use the tools required for developing my PhD research. I would like to express my most profound gratitude to Iztok Kosem for his guidance and support.

## Competing Interests

The author has no competing interests to declare.

## References

- Erlandsen, J.** (2010). Computational lexicography and lexicology iLEX, a general system for traditional dictionaries on paper and adaptive electronic lexical resources. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV EURALEX International Congress* (p. 306). Leeuwarden/Ljouwert: Fryske Akademy – Afûk.
- Gantar, P., Kosem, I., & Krek, S.** (2016). Discovering automated lexicography: The case of the Slovene lexical database. *International Journal of Lexicography*, 29(2), 200–225. DOI: <https://doi.org/10.1093/ijl/ecw014>
- Hartmann, R. R. K., & James, G.** (1998). *Dictionary of lexicography*. London and New York: Routledge. DOI: <https://doi.org/10.4324/9780203159040>
- Hyland, K.** (2009). *Academic discourse*. London: Continuum International Publishing Group.
- Kilgarriff, A., Baisa, V., Rychlý, P., & Jakubíček, M.** (2015). Longest-commonest match. In I. Kosem, M. Jakubíček, J. Kallas, & S. Krek (Eds.), *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference* (pp. 397–404). Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.
- Kilgarriff, A., Husák, M., Mcadam, K., Rundell, M., & Rychlý, P.** (2008). GDEX: automatically finding good dictionary examples in a corpus. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the 13th EURALEX International Congress* (pp. 425–432). Barcelona: Universitat Pompeu Fabra.
- Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D.** (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 105–115). Lorient : Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.
- Rundell, M., & Kilgarriff, A.** (2011). Automating the creation of dictionaries: where will it all end? In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 257–281). Amsterdam: John Benjamins. DOI: <https://doi.org/10.1075/scl.45.15run>


**How to cite this article:** Kuhn, T. Z. (2019). A Design Proposal of an Online Corpus-Driven Dictionary of Portuguese for University Students. *Journal of Portuguese Linguistics*, 18: 3, pp. 1–4. DOI: <https://doi.org/10.5334/jpl.209>

**Submitted:** 05 October 2018

**Accepted:** 27 December 2018

**Published:** 29 January 2019

**Copyright:** © 2019 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Journal of Portuguese Linguistics* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 