A BERT-Powered Writing Assistant for Academic Purposes in European Portuguese

Sílvia Araújo[1][0000-0003-4321-4511] , Micaela Aguiar[1][0000-0002-5923-9257] and José Monteiro[1]

[1] Universidade do Minho, Braga, Portugal
saraujo@elach.uminho.pt, maguiar60@gmail.com,
jdiogoxmonteiro@gmail.com

Abstract. In this paper, we will present the process of developing a resource that we consider to be useful for both native and non-native college students in the process of writing Portuguese academic texts: a BERT-powered Writing Assistant for academic purposes in European Portuguese. The Writing Assistant includes two main components: a phrase bank, that will be created using open scientific data in the form of scientific papers found in repositories, and a search engine, that uses BERT models for semantic searches. To create the phrase bank we will loosely follow the methodology developed by John Morley, creator of the Academic Phrasebank of the University of Manchester. The phrase bank will be based on 40 scientific papers taken from the repository of University of Minho. The corpus will be initially annotated, using some of the categories proposed by Morley, then the categories will be revised to better represent the reality of Portuguese academic discourse. The annotated phrases will then be simplified and stripped of any particular academic content. This phrase bank will "feed" the search engine. The search engine works with BERT machine learning models that allow us to make semantic searches. Students would just have to write a word, expression or sentence in the search bar to find equivalent or similar expressions on our phrasebank, even if the user has little to no knowledge of the vocabulary used in academic discourse, because Bert models are able to infer semantic context and find relevant results.

Keywords: Academic Literacy, Search Engine, Phrase bank, BERT.

1 Introduction

Students (especially college students) struggle to write in academic contexts. This has been well documented by researchers [1] and teachers [2], [3], and certainly felt by the students themselves. In this paper, we will give an account of the process of developing a BERT-powered Writing Assistant for academic purposes in European Portuguese, an interactive phrase bank that aims to help native and non-native college students in Portuguese academic contexts. We will describe the process of creating the phrase bank for academic European Portuguese and then proceed to take a closer look at the technology that will power the Writing Assistant.

1.1 The problem

Academic literacy is rarely explicitly taught, and institutions struggle with changing demographics, linguistic diversity [4] and an increase in students pursuing postgraduate and doctoral studies [5]. The lack of academic literacy in higher education is usually addressed through (more times than not, paid) academic writing courses, typically too generic and superficial in nature to truly be useful. There is a consensus that teaching academic literacy should start much earlier, before students even reach college [6]; however, this is a policy and educational reform issue that is not easily put into practice. Meanwhile, some students do manage to master academic writing, but for most of them not having enough academic literacy ends up being another inequality factor [7], because it is essential to know how to communicate in academic contexts to participate and be a member of the academic community [8].

1.2 A Project geared towards Open Education

With this in mind, we are developing as part of PortLinguE, a portuguese project financed by European funds, a tool that will assist students in the writing process within academic contexts. This project is geared towards open education and open science, from the perspective of open educational resources, as it uses open data to create new resources, and from the perspective of open educational practices, as we seek to "respect and empower learners as co-producers on their lifelong learning paths" [9] by promoting open learning architectures and citizen science. Indeed, open science practices promote a culture of education and scientific literacy [10].

1.3 The Goals

The starting questions that prompted this work were: How can open scientific data be used to help students develop their academic literacy? And how can we accomplish this goal in an engaging way? So, we set out to create a tool that would help students write in academic contexts through an interactive interface which allows them to search phrases and find similar expressions. Bearing this in mind, we conceptualize a tool that we are currently developing. In this paper, we will give an account of the development process of the Writing Assistant.

Our Writing Assistant will be the first of its kind in European Portuguese: not only the first phrase bank for the academic European Portuguese that uses scientific open data, but also the first to be powered by a search engine. We believe that our Writing Assistant will be a useful tool for both native college students and for Portuguese as Foreign Language students, given the challenges they face when it comes to academic writing.

We know, of course, that academic literacy does not stop at writing conventions, and that is why the Writing Assistant is part of a larger set of resources we are developing to aid academic literacy. For instance, a seven step method that guides students through the process of preparing an academic paper, via informational texts, videos and infographics, is also in preparation.

2 The Writing Assistant — A Methodology

The Writing Assistant includes two main components: a phrase bank that will be created using open scientific data in the form of scientific papers found in repositories, and a search engine that uses BERT models for semantic searches. The next sections present the overall process of developing the phrasebank and the search engine.

2.1 The Phrase Bank

Formulaic Language and Academic Writing. Native speakers favor formulaic language in their communication [11] and, in academic writing, "the absence of such formulaic sequences" may actually be a signal of "lack of mastery of a novice writer in a specific disciplinary community" [12]. However, the use of academic formulas "is not part of the native writer's innate language ability and is thus far from being a linguistic universal skill" [13]. In the same way, formulaic nature is the biggest barrier for L2 learners to sound native-like [11]. That is why an academic phrase bank can be useful for both native-speakers and L2 learners.

A phrase bank is a set of expressions, commonly used in academic writing to perform certain language acts, such as referring to sources, describing the results of an experiment or stating the conclusions of a study.

There are some academic phrase bank available online, such as the Ref-N-Write Academic Phrasebank (https://www.ref-n-write.com/academic-phrasebank/) for English or the Dictionnaire des expressions from Base ARTES (https://artes.app.univ-paris-diderot.fr/artes-symfony/web/app.php/fr) for French. In Portuguese, Bab.la offers an English-Brazilian Portuguese phrase bank (https://pt.bab.la/frases/academico/indice/portugues-ingles). However, the most popular phrase bank is Manchester University's Academic Phrasebank (https://www.phrasebank.manchester.ac.uk), developed by Morley [14].

Morley drew on the concept of 'move' [15] as a section of text serving a particular communicative purpose to organize the phrase bank into its multiple categories and subcategories. According to Morley, the Academic Phrasebank corpus originally consisted of 100 postgraduate dissertations from the University of Manchester, and has since incorporated academic material from a variety of sources. The original phrases were simplified and any particular academic content was removed or replaced so that these expressions could be used freely by students without the risk of plagiarism.

Our work departs from Swales' socio-rhetorical approach and falls into an enunciative-pragmatic perspective of Discourse Analysis [16], [17]. In this framework, we consider the concept of discursive genre [18], [19], and, specifically, that of the "scientific article" genre as essential to identify and categorise speech acts, discourse markers and other phraseological units that occur in academic discourse. Fig. 1 shows the workflow for the creation of our phrase bank:

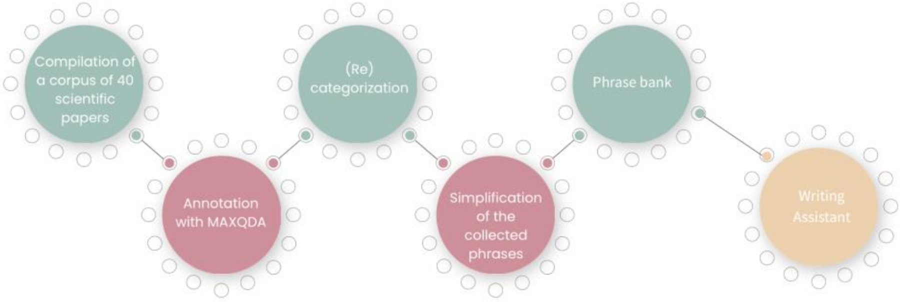
Compilation of a corpus of 40 scientific papers
Annotation with MAXQDA
(Re) categorization
Simplification of the collected phrases
Phrase bank
Writing Assistant

Fig. 1. Workflow for the Creation of the European Portuguese Phrase bank

Compilation. The phrase bank we are developing is based on an initial corpus of 40 scientific articles, taken from RepositoriUM, the repository of the University of Minho, and is divided into four scientific areas, as determined by the Foundation for Science and Technology (FCT): Life and Health Sciences, Exact and Engineering Sciences, Natural and Environmental Sciences, and Social Sciences and Humanities. This division ensures the diversity of the textual materializations [20] from different areas of the academic discourse. Subsequently, we would like to include sources from other genres of academic discourse, such as, for example, dissertations and book reviews. Articles were only included in the corpus if they were written in European Portuguese and were available in open access.

Annotation. The annotation is being carried out using the qualitative data analysis and mixed methods software, MAXQDA. We are using an enunciative-pragmatic approach, that means, we will be annotating speech acts, discourse markers and phraseological units typical in scientific and academic discourse. As a starting point, we will build on the categories determined by Morley [13] and we will consider the questions that the author defined for the inclusion of a given expression in the phrase bank: does it serve a useful communicative purpose in academic text?; does it contain collocational and/or formulaic elements?; are the content words (nouns, verbs, adjectives) generic in nature?; does the combination 'sound natural' to a native speaker or writer of English?

Categorization. After the initial annotation, we will review and refine the categories since they were originally created as a result of analysing a different academic genre (postgraduate dissertations) in a different language. This step is essential in order to better account for the reality of academic discourse in scientific articles written in European Portuguese.

Simplification. After determining the categories, we will perform the extraction of the phrases and move on to the phrase simplification phase. In this phase, the phrases will be stripped of any particular academic content. The result of the categorization and simplification process will be the phrase bank that will “feed” the writing assistant.

Using the writing assistant, students would just have to write a word, expression or sentence in the search bar to find equivalent or similar expressions. The next section will give a brief overview of the technology that will power our writing assistant.

2.2 The Search Engine

Interface. The Writing Assistant we are developing differs from other static phrase banks, due to its interface and search engine, which makes the phrase bank dynamic and interactive. The interface will have an intuitive and user-friendly design, similar to other search engines, such as Google: the user enters a term, an expression or a phrase in the search bar, and the results will be similar expressions.

Keyword-based search vs. Semantic search. Phrase banks with similar interfaces exist, see for example Ref-N-Write Academic Phrasebank for English. However, our Writing Assistant will be unlike these phrase banks, because of the type of search it performs. Generally, other tools of this kind perform “keyword-based” searches in their phrase banks: that is, when the user enters a term, an expression or a phrase, they will get only templates that contain the words present in the user’s search. For example, if we searched for the expression “in conclusion”, the results would be restricted to the lemma of the word "conclusion": we would get results with the word in the singular form, such as “in conclusion, the results show that”, and in the plural form, such as “the most important conclusions are”.

The Writing Assistant that we are developing will perform another kind of search: semantic search. This type of search is unique given its use of machine learning models capable of inferring the semantic context of the templates from the phrase bank and from the user's queries to present relevant templates as results, even if they do not necessarily contain words present in the user's query. For example, if we searched for the expression “in conclusion”, the results could contain expressions such as “in conclusion, the most pressing aspects identified in this study were”, “ we concluded that x has a significant effect on y”, but also results such as “In the analysis of x, it was found that y can effectively replace z” or “In this study, x was not associated with the emergence/aggravation of y”, which are commonly used in the writing of conclusions in scientific papers, even though they do not include the term “conclusion”.

BERT model. The Writing Assistant search engine works with a BERT machine learning model [21]. BERT is a Natural Language Processing model that analyzes text corpus in terms of similarities at word, collocation and sentence level and distributes the processed data based on semantic similarity, thus generating semantic vectors. Below (Fig. 2), we describe the BERT Model Workflow:

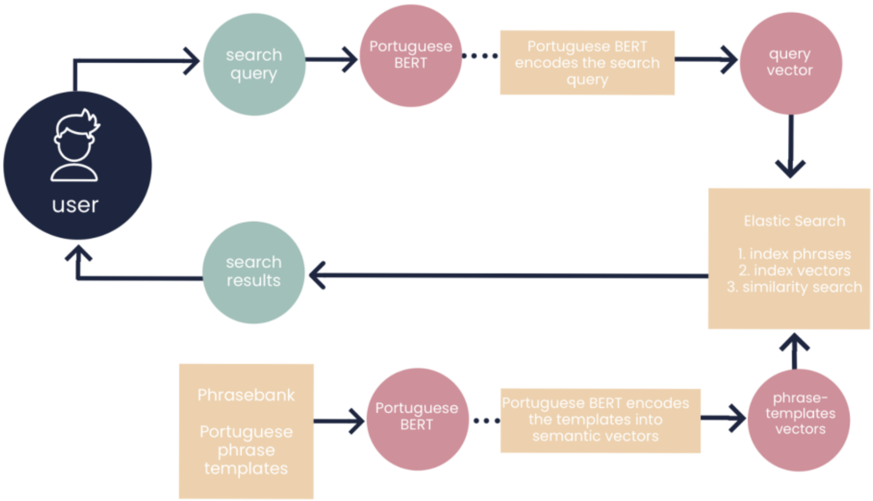

Fig. 2. BERT Model Workflow

Our model will be pre-trained with a Portuguese corpus, so we opted for an open-source model available at BERTinbau, a repository of pre-trained BERT models in Portuguese. This model will process the templates in our phrase bank and the queries entered by the user and it will create the corresponding semantic vectors as a result. The phrase-templates, as well as their semantic vectors, will be stored in an ElasticSearch database. After processing the queries and all the template-phrases, semantic similarity mechanisms will be applied [22]. At the end of this process, the search engine will provide users with the most relevant results. As the vectors generated by Bert models are able to infer semantic context, the engine will be able to find relevant results even if the user has little to no knowledge of the specific vocabulary used in academic discourse.

3 Conclusions & Future Work

This paper set out to give an overview of the development process of the Writing Assistant, a tool that is being created to help students write in academic contexts, as part of a larger project of academic and scientific resources. We looked at the two main components of the Writing Assistant — the phrase bank and the technology — and outlined the different steps that their creation entails. We have presented the workflow behind the creation of the European Portuguese phrase bank: compilation, annotation, categorization and simplification. We have also highlighted the advantages of using semantic search as opposed to keyword-based search and of using BERT models to do so. We have outlined the BERT model workflow starting from the user's search query up to the search result.

Moving forward, the focus will be in testing the Writing Assistant for usability and accuracy when it comes to the technological component. We will also be testing the Writing Assistant for its effectiveness in improving academic writing among native and non-native students. Furthermore, we would also like to build on the original phrase bank to explore how to enrich the corpus with a diverse array of academic genres, using a corpus linguistics approach.

Acknowledgments

This work was carried out within the scope of the "PortLinguE" project (PTDC/ LLT-LIG/31113/2017) financed by FEDER under Portugal 2020 and by national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT,I.P.).

References

1. Defazio, J., Jones, J., Tennant, F., & Hook, S. A.: Academic literacy: The importance and impact of writing across the curriculum – a case study. Journal of the Scholarship of Teaching and Learning 10(2), 34-47 (2010).
2. Estrela, A., Sousa, O.: Competência textual à entrada no Ensino Superior. Revista de Estudos da Linguagem, 19(1), 247–267 (2011).
3. Brandão, J. A.: Literacia Académica: Da Escola Básica ao Ensino Superior — uma visão integradora. Letras & Letras, (2013).
4. Purser, E. R., Skillen, J., Deane, M., Donohue, J., & Peake, K.: Developing academic literacy in context (2008).
5. Ferrão Tavares, C., Pereira, A. L.: Apresentação: Literacias académicas multimodais. Intercompreensão, 16, 5–10 (2012).
6. Gouveia, C. M.: Como se faz uma disciplina: Mapas de conhecimento e distinções operacionais sobre o que é o discurso académico enquanto objeto de estudo. In: Caels, F., Barbeiro, L. F. & Santos, J. V. (eds.) Discurso Académico: Uma Área Disciplinar em Construção, pp. 19–43. CELGA-ILTEC, Universidade de Coimbra/ Escola Superior de Educação e Ciências Sociais, Politécnico de Coimbra (2019).
7. Preto-Bay, A. M.: The Social-Cultural Dimension of Academic Literacy Development and the Explicit Teaching of Genres as Community Heuristics. The Reading Matrix 4 (3) 86–117 (2004).
8. Neeley, S. D.: Academic literacy. Addison Wesley Longman (2005).
9. Ehlers, U.D.: Extending the Territory: From Open Educational Resources to Open Educational Practices. Journal of Open, Flexible, and Distance Learning 15(2), 1-10 (2011).
10. Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., Hartgerink, C. H.: The Academic, Economic and Societal Impacts of Open Access: An Evidence-Based Review', F1000Research 5 (632), (2016).

11. Wray, A.: Formulaic language and the lexicon. Cambridge: Cambridge University Press (2002).
12. Li, J., Schmitt, N.: The acquisition of lexical phrases in academic writing: a longitudinal case study. Journal of Second Language Writing, 18(2), 85–10 (2009).
13. Pérez-Llantada, C.: Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. Journal of English for Academic Purposes, 14, 84–94 (2014).
14. Academic Phrasebank, https://www.phrasebank.manchester.ac.uk, last accessed 2022/04/12.
15. Swales, J.: Genre analysis: English in academic and research settings. Cambridge: Cambridge University Press (1990).
16. Adam, J.-M.: Discursivité, généricité et textualité. Distinguer pour penser la complexité des faits de discours. Recherches, 56, 9-27 (2012).
17. Charaudeau, P.: La situation de communication comme fondatrice d'un genre: la controverse. In : Monte, M. & Philippe, G. (eds.), Genres et textes : déterminations, évolutions, confrontations, pp. 49-57. Lyon, Presses universitaires de Lyon (2015).
18. Bakthine, M.: Esthétique de la Création Verbale. Paris, Gallimard (1984).
19. Maingueneau, D.: Discours et analyse du discours. Paris, A. Colin (2014).
20. Coutinho, M. A.: Descrever géneros de texto: resistências e estratégias. In: Proceedings of the IVth International Symposium on Genre Studies (SIGET), pp. 639–647. Tubarão, Santa Catarina, (2007).
21. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL (2019).
22. Varun. Calculating Document Similarities using BERT, word2vec, and other models. Towards Data Science (2020).