9th International Conference on Higher Education Advances (HEAd'23)
Universitat Politècnica de València, València, 2023
DOI: http://dx.doi.org/10.4995/HEAd23.2023.16314

Data-driven project-based learning in specialized translation classes – the case of comparable corpora

Katrin Herget[1], **Teresa Alegre**[2]

[1]Department of Languages and Cultures / Centre for Languages, Literatures and Cultures, University of Aveiro, Portugal, [2]Department of Languages and Cultures / Centre for Languages, Literatures and Cultures, University of Aveiro, Portugal.

Abstract

The aim of this study is to contribute to data-driven project-based learning in Translation Studies in the context of Higher Education. Based on the implementation of a corpus-based project, a group of second year Master's students in Specialized Translation was actively involved in the compilation and exploration of comparable text corpora in vulgarization of medical science. Comparable corpora are becoming increasingly significant in the field of Language Technology and represent an essential basis for the extraction of terminological and phraseological units. The study falls within the scope of medical eponyms in English and Portuguese corpora. In this context, special attention was paid to the importance of corpus building for the dissemination of medical information to the general public. Against this backdrop, the project enhanced proactive learning by familiarizing students with corpus methodologies, textual conventions, style and register. By compiling and analyzing comparable corpora, students became autonomous researchers and developed skills and competencies for their future professional practice.

Keywords: *Data-driven project-based learning; specialized translation; comparable corpora; medical eponyms; Portuguese-English.*

Editorial Universitat Politècnica de València

1. Introduction

The aim of this study is to present a data-driven project-based learning approach based on the implementation of a classroom project in a higher education setting. The idea of adopting a project-based learning (PBL) approach is to provide students with real-life work environment, introducing them to a more dynamic and motivating learning setting. We explore how PBL can contribute to students' professional preparation and show how data-driven work can add value to language learning.

Data-driven learning has been widely investigated in the field of Language Pedagogy and Learning. The term is commonly used to describe "tools and techniques of corpus linguistics for pedagogical purposes" (Gilquin & Granger, 2010, p. 359). According to the authors, data-driven learning provides several advantages, such as confronting the learner with authentic language, as well as giving them access to "a large number of authentic instances of a particular linguistic item" (p. 359). In this respect, Lenko-Szymanska & Boulton (2015) claim that

> *learners can benefit tremendously from the direct use of corpora. They gain access to authentic language, which they can query in a variety of ways for the information which is interesting and relevant to them at a particular moment and which allows them to refine their understanding of how language really behaves (p. 3).*

The present study aims at applying a data-driven learning approach within the scope of a translation project on the construction and exploration of comparable corpora, and intends to contribute to the implementation of innovative learning resources in Higher Education.

2. Project-based learning

Over the last decades, there has been an increasing interest in the study on project-based learning (PBL) in higher education settings. According to Thomas (2000, p. 1) "Project-based learning (PBL) is a model that organizes learning around projects" that focus on "questions or problems that 'drive' students to encounter [...] the central concepts and principles of a discipline" (p. 3). This model came up as a response to the need of engaging students in authentic activities, reflecting real-world challenges and needs. Higher education institutions must be aware of these changes and come up with adequate solutions in course design. With respect to this, Uden & Beaumont (2006) believe that "[u]niversity education should, ideally, provide students with the necessary skills, values, and attitudes that are essential to cope with the dynamic complexities of the modern world. […] there is a lack of deep learning about the complex issues and problems that graduates have to face in the real world” (p. 26). Due to a terminological fuzziness, PBL is herein used an umbrella term for innovative teaching activities that center on learning through projects. In the present study, data-driven learning is implemented in the broader context of PBL and provides a specific focus on corpus-guided decision-making / construction and exploration.

3. Comparable corpora in the translation classroom

Due to the importance of Specialized Translation in today’s globalized world, translation classes have to respond to a diversified professional reality, preparing students to actively deal with different communicative situations. According to López-Rodríguez & Tercedor Sánchez (2008), “getting familiar with corpora and annotation” ranks among learner-centered activities that help to develop and promote the learners’ autonomy. In Applied Linguistics, the collection of textual data has long been fundamental for the study of specific lexico-grammatical phenomena and patterns, in order to obtain a better understanding of different language registers and varieties. In the field of Translation Studies, Baker (1993) already predicted, in her seminal work on Corpus Linguistics, "that the availability of corpora and of corpus-driven methodology will soon provide valuable insights in the applied branch of translation studies" (p. 242). Since then, research on corpora in Translation Studies has developed at a fast pace and has become fundamental both in theoretical and in applied studies. Molés-Cases & Oster (2015, p. 204) present a detailed overview of practical fields of application of corpus-driven work in translation training. According to Krüger (2012, pp. 507-508) corpora “allow for a better contextualisation and control of the texts to be investigated and provide a higher representativeness [sic], generalisability [sic] and replicability of the findings”. Depending on the research field and objectives, different types of corpora are used. Translation Studies mainly distinguish between comparable and parallel corpora. Due to advances in Machine Translation, the importance of collecting comparable data is getting more and more significant. Comparable corpora are an essential basis for the extraction of bilingual dictionaries, because, unlike parallel corpora, there is no influence from the source text structure. According to McEnery & Hardie (2012, p. 20), a comparable corpus contains “components that are collected using the same sampling method, e.g. the same proportions of the texts of the *same genres* in the *same domains* in a range of different languages in the *same sampling period*” [emphasis in the original]. In this sense, Mikhailov & Cooper (2016, p. 217) also define comparable corpora as text collections that were compiled “on the same principles (size of the collections, size of the samples, topics covered, chronological period, etc.) in different languages, or different variants of the same language: e.g. texts on atomic energy in French and Spanish, or texts in the German of Germany, Austria and Switzerland”. According to Bernardini, Stewart & Zanettin (2003, p. 6) comparable bilingual corpora are important for translation students in that they provide them with an understanding of both target and source texts, "allowing them to compare terminology, phraseology and textual conventions across languages and cultures". In the field of Specialized Translation comparable corpora assume a fundamental role as they provide evidence of specific lexico-grammatical structures, as well as information on the frequency of linguistic patterns in specialized domain texts.

4. Motivation for the Study

In times of unlimited access to information through the internet in form of online dictionaries, databases, or machine translation systems, translation students need to be aware of the challenges arising from the amount of available data. In order to choose suitable texts for the compilation of ad-hoc corpora in the field of medical science vulgarization, students are required to assess the quality of online-texts according to the purpose of the task and to apply a set of criteria to ensure the usability and reliability of the respective data. Against this backdrop students were familiarized with the advantages of corpus compilation and exploration for their future professional work. Another motivation for the study results from the fact that the source language of a text available on the internet is not always obvious, due to a growing amount of machine-translated texts. Therefore, it is a fundamental requirement to make sure that the source text is authentic and that it is not the result of a translation.

5. Method and Study Design

The aim of this study is to actively involve a group of 12 MA students of Specialized Translation in the compilation and exploration of comparable text corpora in the field of vulgarization of medical science in Portuguese and English. The corpus analysis was carried out with the help of the concordance program AntConc (Anthony, 2022). The project was subdivided into three stages, which will be presented in the following.

Stage 1 - Project preparation: Contextualization and pre-corpus building stage

In this preliminary stage, students learn to search and select websites from the field of medical sciences in Portuguese and English by using a specific set of criteria: i) websites belong to the same domain; ii) texts belong to the same genre; iii) texts aim to address a general public on health issues; iv) texts are written by health specialists or technical journalists; v) texts belong to the same sampling period, and vi) texts are written in English / European Portuguese.

1.1 Corpora in specialized translation

a) Identifying different types of corpora (monolingual, bilingual, multilingual, parallel, and comparable) with the main focus being on comparable corpora;

b) Reflecting on the advantages of comparable corpora (e.g. comparable corpora give evidence of terminology and textual conventions in a particular language context);

c) Importance of ad-hoc corpora for Specialized Translation (e.g. compilation of texts for the analysis of specific lexico-grammatical patterns, language varieties, communication levels, etc.);

d) How comparable are comparable corpora? (comparability in terms of dimension, field area, subject content, target audience, sender-receiver relationship, time period, textual genres, mode, etc.).

1.2 Medical eponyms and their use in popularization texts

a) Identification of medical eponyms (e.g. presentation of Wordcloud, Figure 1, as an inductive learning approach, etc.);

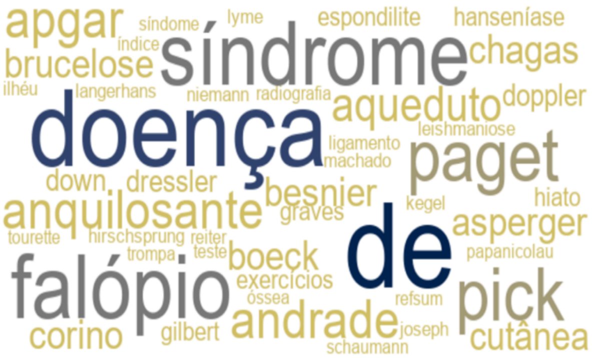


Figure 1: Wordcloud representing the frequency of eponyms in the Portuguese corpus.

b) Variability of eponyms in medical vulgarization texts (synonyms, concurrent designations, etc.);

c) Identification of specific translation problems related to medical eponyms (e.g. an eponym (*Paget*) denotes two different diseases: bones versus breast cancer).

Stage 2 - Project execution

2.1 Compilation of ad-hoc corpora

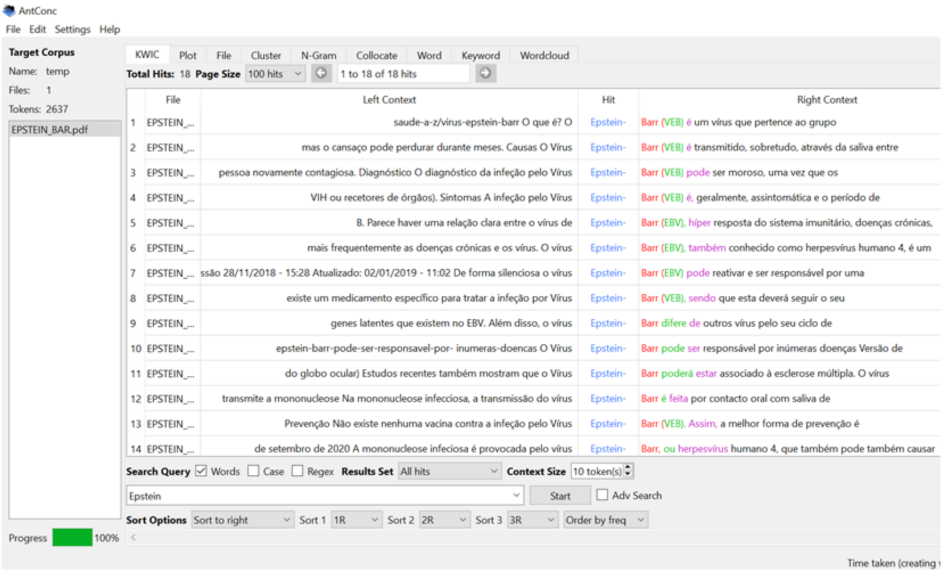
a) Selection of adequate websites (vulgarization of medical science);

b) Discussion on the adequacy and reliability of the findings.

2.2 Corpus compilation and concordance analysis with AntConc

a) Compilation of reliable texts according to criteria such as: domain, authorship, target-audience, language variety, chronological period, etc.;

b) Concordance search and extraction of medical eponyms (Figure 2);



Target Corpus
Name: temp
Files: 1
Tokens: 2637
EPSTEIN_BAR.pdf

Total Hits: 18 Page Size 100 hits 1 to 18 of 18 hits

	File	Left Context	Hit	Right Context
1	EPSTEIN_...	saude-a-z/virus-epstein-barr O que é? O	Epstein-	Barr (VEB) é um vírus que pertence ao grupo
2	EPSTEIN_...	mas o cansaço pode perdurar durante meses. Causas O Vírus	Epstein-	Barr (VEB) é transmitido, sobretudo, através da saliva entre
3	EPSTEIN_...	pessoa novamente contagiosa. Diagnóstico O diagnóstico da infeção pelo Vírus	Epstein-	Barr (VEB) pode ser moroso, uma vez que os
4	EPSTEIN_...	VIH ou recetores de órgãos). Sintomas A infeção pelo Vírus	Epstein-	Barr (VEB) é, geralmente, assintomática e o período de
5	EPSTEIN_...	B. Parece haver uma relação clara entre o vírus de	Epstein-	Barr (EBV), híper resposta do sistema imunitário, doenças crónicas,
6	EPSTEIN_...	mais frequentemente as doenças crónicas e os vírus. O vírus	Epstein-	Barr (EBV), também conhecido como herpesvírus humano 4, é um
7	EPSTEIN_...	ssão 28/11/2018 - 15:28 Atualizado: 02/01/2019 - 11:02 De forma silenciosa o vírus	Epstein-	Barr (EBV) pode reativar e ser responsável por uma
8	EPSTEIN_...	existe um medicamento específico para tratar a infeção por Vírus	Epstein-	Barr (VEB), sendo que esta deverá seguir o seu
9	EPSTEIN_...	genes latentes que existem no EBV. Além disso, o vírus	Epstein-	Barr difere de outros vírus pelo seu ciclo de
10	EPSTEIN_...	epstein-barr-pode-ser-responsavel-por- inumeras-doencas O Vírus	Epstein-	Barr pode ser responsável por inúmeras doenças Versão de
11	EPSTEIN_...	do globo ocular) Estudos recentes também mostram que o Vírus	Epstein-	Barr poderá estar associado à esclerose múltipla. O vírus
12	EPSTEIN_...	transmite a mononucleose Na mononucleose infecciosa, a transmissão do vírus	Epstein-	Barr é feita por contacto oral com saliva de
13	EPSTEIN_...	Prevenção Não existe nenhuma vacina contra a infeção pelo Vírus	Epstein-	Barr (VEB). Assim, a melhor forma de prevenção é
14	EPSTEIN_...	de setembro de 2020 A mononucleose infeciosa é provocada pelo vírus	Epstein-	Barr, ou herpesvírus humano 4, que também pode também causar

Search Query ☑ Words □ Case □ Regex Results Set All hits Context Size 10 token(s)
Epstein □ Adv Search
Sort Options Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by freq

Figure 2: Results of concordance search on Epstein-Barr in AntConc.

c) Results discussion and group reflection on the variation of eponyms in both corpora.

Stage 3 - Evaluation
a) Evaluation of the project by means of a portfolio, including results from the eponym concordance search, extraction of relevant text segments and respective analysis, as well as reflection on the importance of comparable corpora in Specialized Translation;
b) Online questionnaire (ongoing) on students' perceptions regarding the data-driven project work.

6. Results and Discussion

This paper explored a data-driven project-based learning approach in Specialized Translation and aimed at familiarizing a group of 12 MA students with the compilation and exploration of comparable corpora in the field of medical science vulgarization. Data-driven learning based on projects actively involved students in corpus-based research related to eponyms in medical science vulgarization texts for a large audience. The enormous amount of available texts on the internet bears several challenges for translation students, who are required to apply a set of specific criteria to ensure the usability and reliability of the respective data. One of these challenges consisted in identifying a specific communicative situation and defining the target audience for corpus building. Another challenge involved the validation of the selected texts as original, non-translated items. The project consisted of three stages that aimed at developing students' research and analytical skills by involving them actively in the process of searching and compiling comparable corpora. The first stage consisted of project preparation, giving students an overview of context and pre-corpus building requisites. The second stage (project execution) was dedicated to the compilation of ad-hoc corpora, as well as concordance search and extraction of eponyms. At the end of this stage, students were actively engaged in discussion and group reflection on the challenges and opportunities of comparable corpora. The third and last stage entailed evaluation by means of a portfolio and a submission of an online questionnaire (ongoing) to assess students' perceptions. The project enhanced proactive learning by familiarizing students with corpus methodologies, textual conventions, style and register. By compiling and analyzing comparable corpora, students became autonomous researchers and developed skills and competencies for their future professional practice. The results of the survey questionnaire will help to identify possible limitations and make subsequent adjustments with regard to the design and implementation of future projects.

Acknowledgments

The authors would like to thank the MA students for their participation in this project.

References

Anthony, L. (2022). AntConc (Version 4.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

Aronson, J. K. (2014). Medical eponyms: taxonomies, natural history, and the evidence. BMJ 2014;349:g7586 doi: 10.1136/bmj.g7586.

Baker, M. (1993). Corpus Linguistics and Translation Studies – Implications and Applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and Technology*. In Honour of John Sinclair, (pp. 233-252). Amsterdam: John Benjamins.

Bernardini, S., Stewart, D., & Zanettin, F. (2003). Corpora in Translator Education: An Introduction. In F. Zanettin, S. Bernardini & D. Stewart (Eds.), *Corpora in Translator Education* (pp. 1-13). Manchester: St. Jerome Publishing.

Gilquin, G., & Granger, S. (2010). How can DDL be used in language teaching? In A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 359-370). Oxon, New York: Routledge.

Krüger, R. (2012). Working with Corpora in the Translation Classroom. *Studies in Second Language Learning and Teaching Department*. SSLLT 2 (4). 505-525.

Laviosa, S. (2002) *Corpus-based Translation Studies: Theory, Findings, Applications*. Amsterdam, New York: Rodopi.

Lenko-Szymanska, A., & Boulton, A. (2015). Data-driven learning in language pedagogy. In A. Lenko-Szymanska & A. Boulton (Eds.), *Multiple Affordances of Language Corpora for Data-driven Learning* (pp. 1-14). Amsterdam: John Benjamins.

López Rodríguez, C.I. (2016). Using corpora in scientific and technical translation training: resources to identify conventionality and promote creativity. *Cad. Trad.*, Florianópolis, v. 36, nº especial 1, p. 88-120. http://dx.doi.org/10.5007/2175-7968.2016v36nesp1p88.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Mikhailov, M. & Cooper, R. (2016). *Corpus Linguistics for Translation and Contrastive Studies. A guide for research*. London, New York: Routledge

Molés-Cases, T., & Oster, U. (2015). Webquests in Translator Training: Introducing Corpus-based Tasks. In A. Lénko-Szymanska, & A. Boulton (Eds.), *Multiple Affordances of Language Corpora for Data-driven Learning* (pp. 199-224). Amsterdam: John Benjamins.

Thomas, J. W. (2000). A Review of Research on Project-Based Learning. Autodesk Foundation. [https://my.pblworks.org/resource/document/a_review_of_research_on_project_based_learning].

Uden, L. & Beaumont, C. (2006). *Technology and problem-based learning*. Hershey, PA, USA: Information Science Publishing.