AELFE·TAPP 2021

General and Specialised Corpora to Raise Linguistic Awareness in a Language Undergoing the Normalisation Process: Academic Writing in Basque

Itziar Gonzalez-Dios[1], Uxoa Iñurrieta[1,2], Igone Zabala[1]
[1]Ixa group, HiTZ center, University of the Basque Country (UPV/EHU), Basque Country
[2]GOI institute, Basque Summer University (UEU) and University of the Basque Country (UPV/EHU)

Abstract

Academic writing is challenging for many university students in any language, but it is especially difficult for students whose instruction language is still on its way to normalisation and has an unstable academic discourse, such as Basque. This paper explains how corpora can be exploited to raise these students' linguistic awareness. To that end, learning objectives are defined, corpora-based exercises are designed, and the difficulties that students overcome are observed. The focus of this paper are students of scientific and technological degrees in the courses on Basque for Academic Purposes, where they are taught how to solve lexical, grammatical, stylistic and register-related doubts. The final aim of the course is that these students become aware of the functional development of Basque, so that they contribute to it in their professional careers.

Keywords: *Basque for academic purposes, corpus studies, language awareness, data-driven learning*

Introduction

Acquiring skills for academic writing is a challenge for university students in any language, even in English, where students need to specifically work on their English for Academic Purposes (EAP). However, this acquisition is even more difficult for university students instructed in Basque, since Basque is still on its way to normalisation, and the language patterns that they acquire from lessons and teaching materials are unstable, evolving, and often far from optimal (Zabala et al., 2011).

In order to help students overcome these difficulties, two courses on Basque for Academic Purposes (BAP) are offered in all university degrees at the University of the Basque Country (UPV/EHU), which mainly focus on writing and oral skills but are also useful to raise awareness of the sociolinguistic situation of Basque.

[1] https://www.ehu.eus/euskara-orria/euskara/ereduzkoa/
[2] http://lexikoarenbehatokia.euskaltzaindia.eus/
[3] http://garaterm-corpusa.ixa.eus/

Corpus-informed approaches to teaching and material design are considered fundamental for EAP (Timmis, 2015; Cheng & Flowerdew, 2018). These approaches are even more necessary in the case of Basque, in order to monitor the development of academic registers. Moreover, data-driven learning based in the monitor corpora of Basque is required in order to make future Basque professionals understand that BAP is dynamic and requires their lifelong learning as well as their responsibility in the development and consolidation of specialised registers (Zabala et al., 2016). The use of corpora makes students become active researchers instead of mere consumers of teaching materials (Götz, 2012), and it has been demonstrated that this approach can be at least as effective as other traditional teaching approaches (Cheng & Flowerdew, 2018). Students learn to solve doubts and correct errors about grammar and style (Feng, 2014; Johns & Waller, 2015; Dolgova & Mueller, 2019), as well as to critically analyse the usage of vocabulary and phraseology (Rashtchi & Mohammadi, 2017; Szudarki, 2018; Wu et al., 2021), and patterns which are characteristic to academic genres (Timmis, 2015). This way, on the one hand, students acquire skills which are essential for their lifelong learning and, on the other hand, when they enter the professional world, they are aware of the functional development of Basque and are able to contribute to it in an active and effective way.

This paper describes how general and specialised corpora are didactically exploited within the courses of BAP at the UPV/EHU, in scientific-technological domains. The aim is to improve language awareness among students at several levels: morphology, syntax, vocabulary, phraseology, as well as styles and registers.

Methods

Many corpora are freely available for query. Within the courses of BAP on which this paper focuses, three of these corpora were mainly used, two of which are general and one is specialised. The general ones were *Ereduzko Prosa Gaur* (EPG)[1] 'Exemplary Prose Today' (25,1 M words from the period 2000-2006), which is constituted by literature books and media texts, and *Lexikoaren Behatokia* (LB)[2] 'Observatory of the Lexicon' (98 M words from the period 2000-2021), which is a monitor corpus yearly updated with media texts. As for specialised corpora, the *Garaterm*[3] academic corpus was used (over 18 M words from the 2010-2021 period), which collects texts written by lecturers, professors and students at the UPV/EHU. Garaterm is also a monitor corpus which includes teaching materials, scientific articles and academic papers, either spontaneous (without linguistic revision) or corrected (with linguistic editing). This means that not only correct patterns can be found in it, but also frequent and not that frequent incorrect patterns.

Before using these corpora with didactic purposes, the learning objectives to be achieved were determined. Some of these objectives were concerned with the students' awareness about the structure of the Basque language, while some others

were related to the students' awareness about language uses in both general and academic settings, as well as in several specialised domains. Since different learning objectives require the use of different types of corpora, as well as different kinds of search results (i.e. statistics or concordances), the corpora to be used for each objective were also specified. The main learning objectives defined were:

- To identify correct and incorrect inflected nouns and verb forms which are considered frequent error sources, by using concordances in the spontaneous and corrected subcorpora in Garaterm.
- To solve grammar doubts by using statistics and concordances in the EPG and LB general corpora.
- To identify co-occurrence patterns such as collocations and lexical bundles by using advanced search tools in the Garaterm, LB and EPG corpora.
- To investigate how different words or synonyms are used depending on the language register or domain of specialisation, by using statistical data from both general and specialised corpora.
- To learn how words acquire important aspects of their meaning based on their context, by searching for different senses of a word codified in general and specialised dictionaries and relating these senses to the concordances obtained from different corpora and subcorpora.
- To search for the distribution and frequency of different variants of a term, by using the comparison interface in the Garaterm corpus and the LB corpus.

Next, with those objectives in mind, exercises were designed. To work on language correctness, students were asked to search for correct and incorrect language examples in the Garaterm corpus: nouns inflected for the indefinite (such as *atomo* 'atom', *molekula* 'molecule', *landare* 'plant', *motor* 'motor'...) and verb forms (such as **erabili daiteken / erabil daitekeen* 'that can be used') which are often incorrectly written.

Grammatical questions, on the other hand, were proposed to students based on real scenarios, e.g. when correcting texts in class they were asked if a certain phrase should be corrected in the text. An example of this is the phrase *etorkizunera begira* 'looking at the future'. Depending on the context, the syntactic structure -ra *begira* 'looking at', or -ri *begira* 'looking at' are optional (*etorkizunera begi*ra vs *etorkizunari begira*) or exclusive (*aurrera begira* vs **aurreari begira* 'looking ahead'). To solve this problem, students were asked to use the corpora and dictionaries.

As for phraseological units (collocations and formulas), students were requested to look for combinations that are often considered calcs in style books, e.g. *urrats% eman* (lit. give steps 'take steps') and *bezala ezagutzen da* (lit. known like 'known as'), along with their correct counterparts (*urrats% egin* lit. do steps 'take steps'; *deritzo*, *esaten zaio* lit. called 'known as'). Then, students had to compare the statistics obtained from different corpora and look the combinations up in dictionaries. After having analysed

the data, students had to conclude if the statistics provided by the corpora were coherent with what they found in dictionaries.

To work on the vocabulary, a wide variety of exercises were performed: i) identifying the concordances of a given lexical element (for example, *ostalari* 'guest, host, innkeeper') in different corpora and linking them to different meanings listed in general (guest or innkeeper) and specialised dictionaries (in this case, *host* in the domain of Ecology, *host rock* in Geology, *host computer* and *host system* in Computer Sciences, and *host* in Microbiology); ii) analyzing the use and frequencies of a certain term and its synonyms (for instance, *higidura* and *mugimendu* 'movement') across different areas of specialisation and, based on the data, identifying the appropriate one for specialised contexts; and iii) analyzing the variants of a term in general and specialised corpora and subcorpora (e.g. *estimazio* vs *zenbatespen* 'estimation'; *taxon* vs *taxoi* 'taxon'; *ordenagailu* vs *konputagailu* 'computer').

After the corpora were presented and explanations on their use were given, students did the exercises in class time. Some of them were done individually and others in pairs or small groups. The lessons were held in computer labs, where students had access to the corpora search interfaces. The problems that students faced when using corpora were identified by using direct observation techniques during lessons.

Results and discussion

One of the results of these corpora-based exercises was unexpected. The first step when querying corpora is to decide whether one wants to look for the lemma or form, and students tend to have great difficulties in understanding and applying both concepts. This was also the case for the students under study. Since Basque is an agglutinative language, words are mostly constituted by inflectional and derivational morphemes, and it was observed that students do not distinguish them.

Working with corpora inevitably makes students reflect on the form and meaning of words and on the different types of morphemes, since looking for forms or lemmas entails very different outputs. Table 1 shows the results of the form vs lemma (the lemma is also a form) search of ordenagailu: looking for lemmas gives more appearances than looking for the forms (961 vs 234). Consequently, the rest of the statistics (except for the size of the subcorpus, 18 377 222 words) are different.

Table 1: Garaterm results for 'ordenagailu' when searching forms (left) or lemmas (right)

	Form:*ordenagailu*	**Lemma**:*ordenagailu*
Appearances	234	961
Size of the subcorpus	18 377 222	18 377 222

Appearances/1 million	12.73	52,29
Number of users	99	201

Besides, it was observed that looking for correct and incorrect examples of a given grammatical pattern in corpora considerably increased the students' awareness about morphological, syntactic and discourse context. For example, in order to find indefinite noun phrases, it is important to take into account that quantifiers such as *zenbait* 'some' and *hainbat* 'some' require indefinite inflection but that number quantifiers can require either definite or indefinite inflection depending on the context. Being able to identify errors done by experts (most of the texts in Garaterm are created by experts) was also found helpful for students to augment their linguistic self-esteem, since they realised that some of their errors were not individual but collective, that is, belonging to the academic community.

Searching for phraseological units made students pay attention to multiword sequences which are considered incorrect calcs in writing guides, even if they are very frequently used by teachers and students, since the development of academic registers is yet far from optimal in Basque. By using spontaneous and controlled corpora, students could verify that, even if stigmatised combinations appear in any kind of corpora, those considered more genuine present a higher frequency.

It was also observed that students found general and specialised corpora useful for two tasks that they constantly need to carry out when they are involved in academic writing, i.e variation management and translanguaging. In fact, university students often use scientific articles and books written in English or other major languages as a reference when writing academic texts in Basque. Sometimes, they use dictionaries and terminological databases to search for the Basque equivalents of specialised terms, and they find more than one denominative variant for the required term. It was noted that students started to use corpora more and more, as a consequence of being urged to explain their decisions when choosing the most appropriate variant for their texts. Furthermore, students often use machine translation systems as an aid for academic writing, and it was discovered that exercises about semantico-pragmatic reflections on lexical items in corpora considerably raised their capacity to critically analyse machine translation output, especially concerning the adequacy of specialised lexical items.

Conclusions

This paper focuses on the methodology followed within the courses of Basque for Academic Purposes to increase the students' language awareness through the use of corpora, Basque being a language which is still under normalisation and whose

functional registers are not completely fixed yet. More specifically, the learning objectives were defined, a set of exercises were designed, and the problems that students had when consulting corpora were observed, including issues with morphology, syntax, vocabulary, phraseology, styles and registers. These corpora-based exercises resulted to be fruitful to make students reflect both on language and on the responsibility that future professionals, including themselves, have in the development and stabilisation of specialised registers in Basque. The exercises were also useful to increase the students' linguistic self-esteem, since they were based on authentic and probabilistic data, and some of their difficulties are not individual but related to the ongoing normalisation process of Basque. Regarding future work, additional exercises are planned which take into account the students' Basque language proficiency, their domain of specialisation and other criteria. It is also planned to produce an exercise book or website compiling exercises for each specialisation domain.

Acknowledgements

This work has been partially funded by the projects HARTAvas (PID2019-109683GB-C22) and DeepReading (RTI2018-096846-B-C21).

References

[1]. Chen, M., & Flowerdew, J. (2018). A critical review of research and practice in data-driven learning (DDL) in the academic writing classroom. *International Journal of Corpus Linguistics*, 23(3), 335-369.

[2]. Dolgova, N., & Mueller, C. (2019). How useful are corpus tools for error correction? Insights from learner data. *Journal of English for Academic Purposes*, 39, 97-108.

[3]. Feng, H.-H. (2014) A pilot study: The use of corpus concordancing for ESL learner self error-correction. *Journal of Interactive Learning Research*, 25, 5–25.

[4]. Götz, S. (2012). Testing task types in data-driven learning: Benefits and limitations. *Aufgaben*, 249-276.

[5]. Rashtchi, M., & Mohammadi, M. A. (2017). Teaching lexical bundles to improve academic writing via tasks: Does the type of input matter?. *Electronic Journal of Foreign Language Teaching*, 14(2).

[6]. Szudarski, P. (2018). *Corpus Linguistics for Vocabulary. A guide for research*. London: Routledge.

[7]. Timmis, I. (2015). *Corpus Linguistics for ELT. Research and Practice*. London: Routledge.

[8]. Wu, S., Fitzgerald, A., Yu, A., & Chen, Z. (2020). What are language learners looking for in a collocation consultation system? Identifying collocation look-up patterns with user query data. *ReCALL*, 1-19.

[9]. Zabala, I., San Martin, I., Lersundi, M., & Elordui, A. (2011). Graduate Teaching of Specialized Registers in a Language in the Normalization Process: Towards a Comprehensive and Interdisciplinary Treatment of Academic Basque. In S. Maruenda Bataller & B. Clavel-Arroitia (eds.), *Multiple Voices in Academic and Professional Discourse* (pp. 208-218). Newcastle, England: Cambridge Scholars Publishing.

[10]. Zabala, I., San Martin, I., & Lersundi, M. (2016). Learning terminology in order to become active agents in the development of Basque biomedical registers. *CercleS*, 6 (1), 145-165.

Multilingual academic and professional communication in a networked world

Proceedings of AELFE-TAPP 2021 (19th AELFE Conference, 2nd TAPP Conference)
ARNÓ, E.; AGUILAR, M.; BORRÀS, J.; MANCHO, G.; MONCADA, B.; TATZL, D. (EDITORS)
Vilanova i la Geltrú (Barcelona), 7-9 July 2021
Universitat Politècnica de Catalunya
ISBN: 978-84-9880-943-5