HARTAes-vas: Lexical combinations for an academic writing aid tool in Spanish and Basque

HARTAes-vas: Combinaciones léxicas para una Herramienta de ayuda a la redacción de textos académicos en español y en vasco

Margarita Alonso-Ramos [1] and Igone Zabala [2]

[1] *Universidade da Coruña and CITIC, Campus da Zapateira s/n, A Coruña, 15071, Spain*
[2] *Universidad del País Vasco/Euskal Herriko Unibertsitatea, Barrio Sarriena s/n, Leioa, 48940, Spain*

Abstract

Academic writing has become a priority object of study especially in English, for which there are already many resources to help novice writers. This is not the case for Spanish university students who do not have many writing aids at their disposal. Here we focus on routinized lexical combinations that characterise academic discourse in Spanish and Basque. The aim is to extract these combinations from two academic corpora in order to build a writing aid tool serving both languages.

Keywords

Academic writing, collocations, discourse functions, writing aid.

1. Introduction

The HARTAes-vas project is funded by the Ministry of Science and Innovation in the 2019 call for R&D Knowledge Generation Projects. It is a project coordinated between the Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU) and the Universidade da Coruña (UDC) and, in some objectives, it is a continuation of previous projects related to academic writing in Spanish. In this new project, we are tackling a contrastive approach with two different languages from both a typological and a sociolinguistic point of view. The research team is made up of members of the LyS group at the UDC and the Ixa group at the UPV/EHU together with researchers from the Foundation Elhuyar.

In recent years, academic writing has become a priority object of study, especially in English ([1], [2] among others). In order for members of the academic community to produce knowledge, they must be able to write in the conventional forms of academic texts. However, when students enter university, they are confronted with new written genres for which they are not provided with tools to facilitate the production of texts. Moreover, university students in Spain must be able to show proficiency in several languages and, paradoxically, Spanish students have more resources to help them with academic English than with the other languages of the state. One of the keys to this competence in writing lies in the mastery of certain routine expressions that give it its specific character: *academic lexical combination* (ALC), ranging from collocations (*extraer conclusiones*, *ondorioak atera* 'draw conclusions'), to discourse markers (*en conclusión*, *ondorioz* 'in conclusion') and also formulas such as *parece razonable concluir que* ('it seems reasonable to conclude that'), *ondorioz esan daiteke* ('consequently we can say'); all

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain
EMAIL: margarita.alonso@udc.es (M. Alonso-Ramos); igone.zabala@ehu.eus (I. Zabala)
ORCID: 0000-0002-1353-9270 (M. Alonso-Ramos); 0000-0002-1931-4136 (I. Zabala)

these expressions are ALC which we have in order to express a conclusion in Spanish and Basque.

Before developing the tool that would help the students learn to write in this academic style, a diagnosis of the current written productions of our university students is needed. In previous research we have compiled a corpus of written productions of Spanish academic novices made up of Bachelor's and Master's theses ([3], [4]; hereafter the Spanish novice corpus) and during this project we have compiled a comparable corpus of written productions of academic novices for Basque (hereafter the Basque novice corpus). The different sociolinguistic status of Basque with respect to Spanish forces different strategies: on the one hand, there is no academic corpus of expert academic writing in Basque available as a reference; on the other hand, Basque has not had enough time for the stabilisation of academic registers [5], which suggests as a starting hypothesis that ALCs will have a lower degree of fixation and recurrence. Likewise, the agglutinative nature of Basque poses a challenge to the usual techniques for extracting combinations.

2. Goals

The overall goal is to create a bilingual tool (or two coordinated monolingual tools), focused on the use of ALCs, combining a dictionary and a corpus. We aim to build a tool where the user can choose the language and find help in choosing the appropriate lexical strategies according to different discourse needs.

More specifically, the project aims to:

- develop a model of ALCs that includes the characteristics of agglutinative languages such as Basque where different lexicographic and discursive classifications will be established;
- analyse the learners' use of such combinations in Spanish and Basque;
- investigate what kind of help related to the phenomena of lexical combinations they need when writing;
- develop corpus-based linguistic technologies for the automatic identification of ALCs.

3. Methodology

The project has multiple orientations: lexicological (as far as the linguistic phenomena studied are concerned); corpus linguistics and computational linguistics (insofar as the corpora are the fundamental source of data and the techniques with which they are exploited come from NLP) and didactics (following the approach of so-called *computer-assisted language learning* and, more particularly, the *data-driven learning* methodology).

The agglutinative nature of Basque inspired the design of alternative ALC identification techniques since the usual lexical bundle extraction technique is not suitable in all cases for Basque. The reason is that some formulas are made up of a single word in Basque and it is necessary to take into account the so-called *morphemic bundles* to complement the results obtained with the techniques used for inflectional languages. For example: *en resumen* 'in short' - *laburbilduz* 'short+gather+INSTR'; *por consiguiente* 'therefore'- *ondorioz* 'consequence + INSTR'.

3.1. Extracting academic vocabulary lists with corpus linguistics and NLP techniques

We analysed the Spanish novice corpus morphologically and syntactically to extract collocations with LinguaKit, Freeling and UDPipe, following the same criteria we used in the expert corpus [4]. We extracted the following syntactic patterns: Subject-Verb (*objetivo se centra* 'objective focuses'), Verb-Object (*alcanzar objetivo* 'reach an objective'), Noun-Modifier (*objetivo fundamental* 'main objective'), N of N (*serie de objetivos* 'series of objectives'). We also extracted lists of n-grams, applying criteria of frequency and distribution by scientific domains and assigned the discursive function according to the typology established in [6].

A similar procedure was applied to the Basque novice corpus which was morphologically analysed using Eustagger. We started by extracting an academic vocabulary based on the criteria defined in [7]. We have used this word list to identify collocations, without the need to syntactically analyse the corpus [8]. We have extracted the following syntactic patterns:

Subject-Verb (*datuek erakutsi* 'data show'), Verb-Object (*datuak bildu* 'collect data', *datuetan oinarritu* 'rely on data'), Noun-Modifier (*datu esanguratsu* 'significant data'), N-N (*datu sorta* 'data set', *datu-bilketa* 'data collection'). To obtain the formulas, we extracted lists of n-grams, applying the same criteria of frequency and dispersion and the same typology of discursive functions described in [6]. Once the formula candidates have been validated, the variation was analysed in order to identify prototypical formulas and their variants.

3.2. Testing distributional semantics strategies

Once the two corpora of Spanish and Basque novice academic writing are balanced, we can exploit them as comparable corpora and apply computational techniques of distributional semantics in order to find correspondences between the formulas of the two languages. With the Spanish list, vector representations (embeddings) of each formula can be generated using non-compositional strategies, and we can then use them to identify the Basque single word equivalents of Spanish expressions in a previously obtained cross-linguistic semantic space. In this way, we may be able to relate *por consiguiente* and *ondorioz*, or *para terminar* 'to conclude' and *bukatzeko*, following the non-compositional strategy used by [9].

Monolingual distributional models, both monolexical and polylexical, will be generated with *fastText*, and mapped to a multilingual space with *vecmap*. Since we find both compositional and non-compositional expressions among the formulas, we will use equivalent search strategies adapted to each type of structure. For the non-compositional ones, we will represent each formula with a single vector, using the non-compositional method presented in [9]. We consider that the use of this multilingual strategy can help in the identification of formulas, because if a Basque expression has a high degree of both internal cohesion and distributional similarity with a Spanish formula, the probability that it is indeed a formula in Basque is also very high. Likewise, it seems interesting to explore whether distributional models also identify a more discursive meaning, such as that of the formulas.

4. Results

The quantitative data from the Spanish novice corpus analysis are shown in Table 1. The data are presented with normalised frequency per million words due to the different size of the corpora.

Table 1
The ALC data from the Spanish novice corpus

ALC	Types/M	Tokens/M
N-modif	192	2724
N *de* N	85	1106
Subject-V	39	313
V-Object	219	2753
Formulas	211	20474

The results of a contrastive analysis with the expert corpus show that novices use fewer collocations than experts. Also, novices use more collocations belonging to the general language. With respect to formulas, we see that novices use fewer types than experts, but almost as many tokens

As far as Basque is concerned, we have already achieved the compilation of a corpus of novice academic writing [10]. Although its analysis has not yet been completed, we can already observe some characteristics: the ALCs are less stable compared to the Spanish novel corpus and a higher number of ALCs are considered incorrect. By validating the lists of ALCs in the Basque corpus, we will be able to make a more thorough comparison: contrasting formulas by functions and verifying whether the same functions are covered in the two languages and checking whether the equivalent bases are linked to more or fewer collocates in the different languages. This comparison will be vital for the design of the writing aid tool. Pending the aforementioned further analysis, the quantitative data are shown in Table 2.

Table 2
The ALC data from the Basque novice corpus

ALC	Types/M	Tokens/M
N-modif	150	4024
N - N	43	1251
Subject-V	3	58
V-Object	108	4136
Formulas	196	38171

5. Conclusions and future work

We have presented the main tasks we carried out to obtain the data for an academic writing aid tool. Next, we will explore the transfer strategies for the automatic identification of ALCs in several languages. We start from the hypothesis that a cross-linguistic language model trained to identify the formulas in Spanish could recognise expressions with similar characteristics in Basque. If the results obtained with this strategy are adequate, we could, on the one hand, automatically obtain new formulas in both languages in other corpora and, on the other hand, identify formulas in Basque that could be mapped to those in Spanish. Pending the results of the experiments with distributional semantics techniques, we are making progress in the design of the tool, which must meet two requirements: 1) provide onomasiological access by discursive function; 2) include a field of warnings where examples will be provided as correction models.

Acknowledgements

This work has been supported by the Xunta de Galicia, through grant ED431C 2020/11, by the CITIC of the UDC through grant ED431G 2019/0 and by the Spanish Ministry of Science and Innovation through projects PID2019-109683GB-C21 and PID2019-109683GB-C22. I would like to thank Olga Zamaraeva for her valuable and constructive suggestions.

References

[1] K. Hyland, P. Shaw (Eds.) The Routledge Handbook of English for Academic Purposes, Routledge, London, 2016.

[2] K. Tusting, S. McCulloch, I. Bhatt, M. Hamilton, D. Barton, Academics Writing: The Dynamics of Knowledge Creation, Routledge, Abingdon, NY, 2019.

[3] M. Alonso-Ramos, M. García-Salido, M. Garcia, Exploiting a corpus to compile a lexical resource for academic writing: Spanish lexical combinations, in: I. Kosem, et al. (Eds.), Electronic Lexicography in the 21st Century. Proceedings of eLex 2017 Conference, Lexical Computing Brno, 2017, pp. 571–586.

[4] M. García-Salido, M., M. Garcia, M. Villayandre, M. Alonso-Ramos, 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), Miyazaki, 2018, pp. 260-265.

[5] I. Zabala, M.J. Aranzabe, I. Aldezabal, Retos actuales del desarrollo y aprendizaje de los registros académicos orales y escritos del euskera, Círculo de Lingüística Aplicada a la Comunicación 88 (2021) 31–50. doi: 10.5209/clac.78295.

[6] M. García-Salido, M. Garcia, M. Alonso-Ramos, Identifying lexical bundles for an academic writing assistant in Spanish, in: G. Corpas Pastor, R. Mitkov (Eds.), Computational and Corpus-Based Phraseology. Europhras 2019, volume 11755 of Lecture Notes in Computer Sciences, Springer, Cham, 2019, pp.144–158. doi: 10.1007/978-3-030-30135-4_11

[7] M. García-Salido, Compiling an Academic Vocabulary List of Spanish. Available at: doi:10.13140/RG.2.2.27681.33123.

[8] A. Gurrutxaga, I. Alegria, Automatic extraction of NV expressions in Basque: Basic issues on cooccurrence techniques, in: Proceedings of the Workshop on Multiword Expressions: from parsing and generation to the real world, Association for Computational Linguistics, Portland, 2011, pp. 2–7.

[9] M. Garcia, M. García-Salido, M. Alonso-Ramos, Weighted compositional vectors for translating collocations using monolingual corpora, in: G. Corpas Pastor, R. Mitkov (Eds.), Computational and Corpus-Based Phraseology. Europhras 2019, volume 11755 of Lecture Notes in Computer Sciences, Springer, Cham, 2019, pp. 113–128. doi: 10.1007/978-3-030-30135-4_9.

[10] M. J. Aranzabe, A. Gurrutxaga, I. Zabala, Compilación del corpus académico de noveles en euskera HARTAvas y su explotación para el estudio de la fraseología académica. Procesamiento del Lenguaje Natural 69 (2022) 95-103.