Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 151-154 recibido 15-12-2023 revisado 20-01-2024 aceptado 27-01-2024

Automatic identification of Spanish academic collocations for an online writing tool

Identificación automática de colocaciones académicas en español para una herramienta en línea de ayuda a la redacción

Eleonora Guzzi
LyS Group, University of La Coruña
eleonora.guzzi@udc.es

Abstract: The present Ph.D. thesis was written by Eleonora Guzzi under the supervision of Dra. Margarita Alonso Ramos (University of La Coruña). The defense was held on the 11th of December 2023 at the University of La Coruña and the members of the committee were José Ignacio Pérez Pascual (University of La Coruña), the president, Moisés Almela Sánchez (University of Murcia), the secretary, and Amália Mendes (University of Lisbon), the vocal. The thesis was awarded an excellent grade Cum Laude, and the international doctorate mention.

Keywords: collocations, academic discourse, lexical complexity, corpus, writing tools

Resumen: La presente tesis doctoral ha sido escrita por Eleonora Guzzi, bajo la supervisión de la doctora Margarita Alonso Ramos (Universidad de la Coruña). La defensa se celebró el 11 de diciembre de 2023, en la Universidad de la Coruña, ante el tribunal formado por el presidente, José Ignacio Pérez Pascual (Universidad de la Coruña), el secretario, Moisés Almela Sánchez (Universidad de Murcia) y la vocal, Amália Mendes (Universidad de Lisboa). La tesis obtuvo la calificación de sobresaliente Cum Laude y la mención internacional.

Palabras clave: colocaciones, discurso académico, complejidad léxica, corpus, herramientas de escritura

1 *Goals and motivation*

The main goal of the present thesis is to contribute to the field of Spanish for Academic Purposes through a research on academic collocations within the areas of Corpus Linguistics, Computational Linguistics and Lexicography.

On the one hand, the choice of academic discourse as the scope of this study has been motivated by the fact that one of the greatest difficulties faced by students is to acquire proficiency in academic writing. This challenge arises especially because an advanced knowledge of specific vocabulary is required, among other genre-related knowledge. On the other hand, the decision to focus on collocations, such as *deep analysis*, *confirm hypothesis* or *draw conclusions,* as object of study has been based on the assumption that this type of expressions, along with other phraseological units, enriches the academic prose and contributes to communicative effectiveness. Moreover, research has proved that a higher collocational competence in academic discourse can be synonymous with greater academic success. However, students typically exhibit limited experience in academic writing and have a little exposure to this type of vocabulary: academic collocations are neither part of implicit everyday language nor explicitly taught as the technical vocabulary. In fact, existing literature has highlighted the insufficient familiarity undergraduate students possess with the prototypical phraseology of academic discourse (Boers & Webb, 2018; Paquot & Granger, 2012) and identified collocations as a persistent and frequent challenge in the written competence of both second language learners and less proficient native speakers. In addition, if we explore the field of Spanish academic writing, scarce studies focus on the identification of academic vocabulary and lexicographic resources related

ISSN 1135-5948 DOI 10.26342/2024-72-11

to phraseological expressions and academic writing.

Therefore, this research aims to address identified gaps in quantitative corpus-based studies on Spanish academic vocabulary by compiling a list of academic collocations and providing a comparison of the collocational complexity of expert and undergraduate writings. Moreover, given the scarcity of resources and lexicographic tools in Spanish as far as academic writing is concerned (Alonso-Ramos et al., 2017; Núñez Cortés & Da Cunha, 2022; Guzzi & Alonso-Ramos, 2023a), one of the purposes of this research is precisely to include the Spanish academic collocations list in a writing aid (HARTA; http://www.dicesp.com:8083; Alonso-Ramos et al. 2017), designed with a corpus-dictionary format and aimed to contain academic multi-word expressions in Spanish (Guzzi et al. 2023).

On the other hand, as a result of the thresholds established in the collocational complexity comparison, an automated evaluation system intended for academic Spanish certification exams, such as EXELEEA (Mendoza, 2015) is proposed. Finally, the study of collocations could have multiple practical applications, which in Spanish have not yet been exploited. Thus, the proposal of this thesis could be applied as a didactic resource for the teaching of academic Spanish in writing centers, as well as for students with Spanish as L2 who access Spanish-speaking universities. It could have also applications in the field of Computational Linguistics and Natural Language Processing: as is well known, collocational resources are especially useful for rule-based generation and translation systems.

2 Outline of the dissertation

This thesis comprises eight chapters. Chapter 1 provides an overview of the research's motivation and main objectives. In Chapter 2, the field of academic discourse is introduced, encompassing a review of approaches, to delineate the study's scope and an explanation of key concepts such as Language for Specific Purposes, specialized languages and scientific discourse. In this chapter, a general description of academic vocabulary is presented, as well as the types of phraseological units approached within this type of discourse, that includes the concept of collocation adopted in this study (Mel'čuk, 2015). Automatic, statistical and manual methods for identifying academic words and collocations are outlined through a comprehensive review of existing vocabulary lists. The final section explores the concept of lexical and phraseological complexity related to vocabulary assessment, involving parameters such as diversity and sophistication (Crossley, 2020; Kyle & Crossley, 2015; Paquot, 2019). Chapter 3 delves into academic tools addressing phraseological units, ranging from academic corpora to online writing aids. In addition, a first version of HARTA is presented in detail together with a usability test of the tool. Finally, the resources focused on vocabulary assessment related to lexical complexity and lexical profile are addressed.

Chapter 4 details the composition and processing of the two corpora that are employed in this research (expert and novice), using NLP techniques, as well as the methodology for automatically extracting collocation candidates. Chapter 5 explains exhaustively the compilation of the list of Spanish academic collocations. Criteria for filtering collocations are presented, including phraseological and interdisciplinary statistical criteria, along with the method followed to validate the list. Chapter 6 contrasts expert and novice use of collocations, by means of collocational complexity of their texts, that includes the parameters of sophistication, known as the property of lexical items that are less common in general language and are more formal or typical of academic discourse, and diversity, known as the index of repetition of lexical items.

Chapter 7 explores practical applications of the Spanish academic collocation list, emphasizing its integration into the HARTA tool, together with quantitative data and improvements based on the results of the usability test. The second part introduces a beta version of an evaluation tool to automatically calculate the collocational profile of texts, aimed for teachers, evaluators, and researchers. Finally, Chapter 8 summarizes the main conclusions, acknowledges limitations, and outlines future directions for research.

3 Main contributions

From this research, four main contributions can be retrieved.

3.1. Spanish Academic Collocations List

The first one has been the development of a reference list of 5.402 Spanish academic collocations based on a large expert corpus, consisting of scientific articles from 12 different disciplines.

This includes a detailed description of the procedure and criteria we followed to identify these collocations: an automatic extraction process and a manual and statistical review that includes phraseological and distribution criteria. The phraseological criteria allowed us to discard either free combinations or idioms. This phase has proved to be the most demanding due to the large number of candidates automatically extracted and because sometimes the boundaries between types of combinations can be blurred. Even this selection process, some specialized collocations still persisted in the selection, that highlighted the need of an interdisciplinarity analysis. It involved distribution filters and the identification of the different senses of the collocations' bases to discard specialized units.

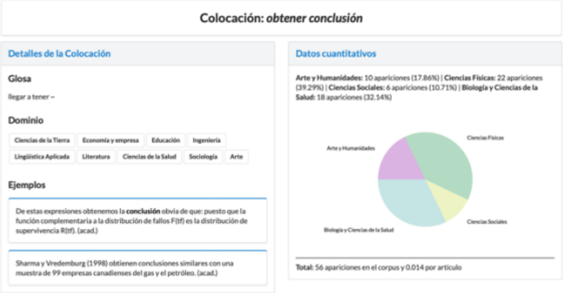
Furthermore, a comparison of the collocations obtained following this method is compared to the collocations that would have been obtained if two association measures (log-likelihood and Mutual Information) were used above a specific threshold. This comparison has shown that the overlapping degree is not elevated but that the log-likelihood measure could be better than Mutual Information for the identification of Spanish academic collocations with a lower threshold.

3.2. Collocational profile of expert and novice writings

The second result concerns the contrastive analysis of the use of academic collocations by experts and novice writers by means of collocational complexity (Guzzi & Alonso-Ramos, 2023b). Results have shown that experts have a wider repertoire of collocations and use those that are stylistically more salient in academic discourse. However, results also suggested that, sometimes, they repeat the same collocation several times in the same text. On the other hand, results have indicated that scientific areas may influence the score of collocational complexity: Biology and Health Science is the field in both groups (expert and novices) where more sophisticated collocations are used, but Social Sciences shows the highest amount of academic collocations. Moreover, a correlation between the linguistic general quality of text and level of collocational complexity was corroborated. Finally, the results obtained allowed us to establish a threshold for scoring academic texts according to the number of collocations, diversity and sophistication. For this purpose, we applied a series of formulas relating to collocational diversity and sophistication in order to obtain the collocational profile of the texts analyzed, understood as an image reflecting the collocational competence of academic texts' writers.

3.3. Improvements of HARTA

The third result is related to the integration of collocations in the writing aid tool HARTA, with improvements in their accessibility, functionality and amount of information. Data about the frequency and distribution of academic collocations has been integrated with a clearer view, as shown in Figure 1, as well as collocations associated to two possible meanings. Those improvements have been implemented as a result of the establishment of the collocation list and the usability test of the tool.



Colocación: *obtener conclusión*

Detalles de la Colocación

Glosa

llegar a tener ~

Dominio

Ciencias de la Tierra | Economía y empresa | Educación | Ingeniería | Lingüística Aplicada | Literatura | Ciencias de la Salud | Sociología | Arte

Ejemplos

De estas expresiones obtenemos la **conclusión** obvia de que: puesto que la función complementaria a la distribución de fallos F(tf) es la distribución de supervivencia R(tf). (acad.)

Sharma y Vredemburg (1998) obtienen conclusiones similares con una muestra de 99 empresas canadienses del gas y el petróleo. (acad.)

Datos cuantitativos

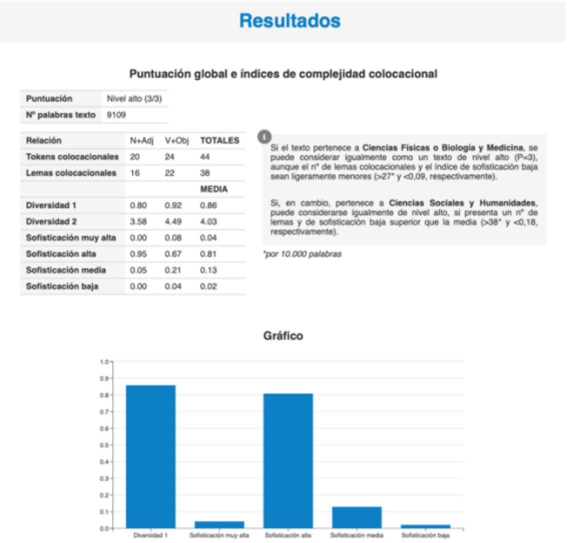
Arte y Humanidades: 10 apariciones (17.86%) | **Ciencias Físicas:** 22 apariciones (39.29%) | **Ciencias Sociales:** 6 apariciones (10.71%) | **Biología y Ciencias de la Salud:** 18 apariciones (32.14%)

Total: 56 apariciones en el corpus y 0.014 por artículo

Figure 1. Lexicographic entry of the academic collocation *obtain conclusion* in HARTA.

3.4. Evaluation tool for the collocational complexity of academic texts

The fourth and last contribution has been the proposal for an automatic evaluation system of the collocational competence of writers through the analysis of collocational complexity. Using a series of Python scripts, a text is processed and a set of indexes from the text are identified that includes: the number of words; the collocational lemmas and their frequency; and the collocations diversity and sophistication, along with a global score, as shown in Figure 2.



Resultados

Puntuación global e índices de complejidad colocacional

Puntuación	Nivel alto (3/3)
Nº palabras texto	9109

Relación	N+Adj	V+Obj	TOTALES
Tokens colocacionales	20	24	44
Lemas colocacionales	16	22	38
			MEDIA
Diversidad 1	0.80	0.92	0.86
Diversidad 2	3.58	4.49	4.03
Sofisticación muy alta	0.00	0.08	0.04
Sofisticación alta	0.95	0.67	0.81
Sofisticación media	0.05	0.21	0.13
Sofisticación baja	0.00	0.04	0.02

Si el texto pertenece a **Ciencias Físicas o Biología y Medicina**, se puede considerar igualmente como un texto de nivel alto (P=3), aunque el nº de lemas colocacionales y el índice de sofisticación baja sean ligeramente menores (>27* y <0,09, respectivamente).

Si, en cambio, pertenece a **Ciencias Sociales y Humanidades**, puede considerarse igualmente de nivel alto, si presenta un nº de lemas y de sofisticación baja superior que la media (>38* y <0,18, respectivamente).

*por 10.000 palabras

Gráfico

Figure 2. Part of the results that the evaluation tool on collocational complexity shows when a text is analyzed.

Acknowledgments

This research was funded by the following grants, centers and institutions: Ministerio de Ciencia e Innovación (PID2019-109683GB-C21); Programa de Axudas á Etapa predoutoral da Xunta de Galicia, FSE Galicia 2014-2020; Centro de Investigación de Galicia do Sistema Universitario de Galicia "CITIC", funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

References

Alonso-Ramos, M., M. García-Salido, and M. Garcia. 2017. Exploiting a corpus to compile a lexical resource for academic writing: Spanish lexical combinations. In Iztok Kosem, Jelena Kallas, Carole Tiberius, Simon Krek, Miloš Jakubíček and Vít Baisa (eds.), *Electronic lexicography in the 21st century, Proceedings of 2017 eLex Conference* (pages 571-586). Brno (Czech Republic).

Boers, F. and S. Webb. 2018. Teaching and learning collocation in adult second and foreign language learning. *Language Teaching*, 51:1, 77-89.

Crossley, S. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11:3, 415-443.

Guzzi, E. and M. Alonso-Ramos. 2022. Selección de colocaciones académicas en español a través de un filtro de interdisciplinariedad. *Procesamiento del Lenguaje Natural*, 69, 83-94.

Guzzi, E. and M. Alonso-Ramos. 2023a. Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, 2, 1-22.

Guzzi, E. and M. Alonso-Ramos. 2023b. Sofisticación y diversidad como medidas de complejidad léxica para determinar el perfil colocacional de textos académicos en español. *Revista Signos*, 56:112.

Guzzi, E., M. Alonso-Ramos, M. Garcia, and M. García-Salido. 2023. Annotation of lexical bundles with discourse functions in a Spanish academic corpus. In Archna Bhatia, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, y Shiva Taslimipoor (Eds.), *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)* (pages 99-105), Association for Computational Linguistics. Dubrovnik (Croatia).

Kyle, K. and S. Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49:4, 757-786.

Mel'čuk, I. 2015. Clichés, an understudied subclass of phrasemes. *Yearbook of Phraseology*, 6:1, 55-86.

Mendoza, A. 2015. La validez en los exámenes de alto impacto: Un enfoque desde la lógica argumentativa. *Perfiles educativos,* 37:149, 169-186.

Paquot, M. and S. Granger. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32, 130-149.

Paquot, M. 2019. The phraseological dimension in interlanguage complexity research. *Second language research,* 35:1, 121-145.