Revista de LINGÜÍSTICA Y LENGUAS APLICADAS

Revista de Lingüística y Lenguas Aplicadas
Vol. 20 año 2025, 32-44
EISSN 1886-6298
https://doi.org/10.4995/rlyla.2025.22588

THE SPANISH ACADEMIC COLLOCATION LIST

Eleonora Guzzi
Universidade da Coruña, Spain

Abstract: *Phraseology plays a crucial role in academic texts, with collocation use key to demonstrating high competence in academic writing. Even for native speakers with limited experience in academic genres, academic collocations present challenges. Despite their importance, however, no sufficiently representative repertoire of academic collocations in Spanish has been developed as a resource for students, similar to those available for academic English. To address this gap, this article proposes a reference list of academic collocations in Spanish, designed for integrated in an academic writing tool and as a support for Spanish for Academic Purposes. Collocations were extracted from an academic corpus using NLP techniques and applying frequency and distribution criteria and Universal Dependency parsing. The resulting list was manually validated to ensure it includes useful collocations for Spanish students across various academic fields.*

Key words: *academic writing, collocations, universal dependencies, phraseology.*

1. INTRODUCTION

One of the most significant challenges for students at university level is acquiring proficiency in academic writing. This difficulty arises from the need to handle multiple textual genres and register conventions, which also requires a mastery of prototypical academic vocabulary. As is well known, academic success is closely tied to linguistic competence, and vocabulary – specifically, phraseological units – is central to the conventionalized nature of academic language (Biber *et al.*, 2004; Hyland, 2008; Paquot, 2012). However, mastering the specialized vocabulary of academic language poses a challenge for all students, irrespective of their native language (Hyland, 2006; Römer & Arbor, 2009; Mauranen *et al.*, 2016), because it is cognitively demanding, is seldom explicitly taught (Coxhead, 2000) and is not accessible through mere exposure to everyday language (Townsend *et al.*, 2012).

Of the various components of academic vocabulary, collocations represent a particular challenge to acquiring written proficiency. Collocations are lexically constrained, and the fact that they may differ across languages makes their acquisition and use additionally difficult (Nesselhauf, 2005; Paquot & Granger, 2012; Wray, 2013; Boers & Webb, 2018). Even native speakers, capable of recognizing collocations, face considerable difficulties in correct usage, particularly in a specialized context like academic writing (Ackermann & Chen, 2013).

The past two decades have seen substantial research into academic collocations, primarily in English, but also in French. In Spanish, however, the literature on academic collocations is notably scarce, with most research paid to other types of phraseological or academic vocabulary (e.g., Errázuriz Cruz, 2014; Yao, 2022) and on academic writing tools focused on grammatical and lexical correctness, but not specifically covering academic collocations, e.g., Estilector (Nazar & Renau, 2023) and arText (Da Cunha *et al.*, 2017).

To address this gap, we propose a reference list of Spanish Academic Collocations (Colocaciones Académicas del Español; CAE) that can serve both as an educational resource within the framework of Spanish for Academic Purposes and a support to lexicographic academic writing tools, such as HARTA (Herramienta de Ayuda a la Redacción de Textos Académicos; Alonso-Ramos *et al.*, 2017; García-Salido *et al.*, 2018). It should be noted that although there is some overlap between academic and general Spanish, the identified collocations in this study are considered as "academic" because they occur very frequently within this genre and are common across various fields, thereby fulfilling the concept of *interdisciplinarity*. We do not assert that general academic collocations should be confined solely to academic writing or to specific fields, as some authors (e.g., Hyland & Tse, 2007; Durrant, 2009, 2016) contend. Instead, in line with Frankenberg (2018), our objective is to identify the expressions

To cite this article: Guzzi, E.(2025). "The Spanish Academic Collocation List". *Revista de Lingüística y Lenguas Aplicadas, 20*, 32-44. https://doi.org/10.4995/rlyla.2025.22588
Correspondence author: eleonora.guzzi@udc.es

Received: 2024-10-09 Accepted: 2025-03-30

that are frequent and essential for composing proficient academic texts in Spanish, while also addressing the well-documented challenges that students face in mastering these expressions compared to more informal or general language.

The article is organized as follows: Section 2 reviews existing research on academic vocabulary; Section 3 outlines the research methodology, including academic corpus compilation and collocation identification; Section 4 presents quantitative and qualitative data on the identified collocations; Section 5 analyses the effectiveness of our methodology; and, finally, Section 6 presents our conclusions.

2. PREVIOUS RESEARCH INTO ACADEMIC VOCABULARY

As Hu and Nation (2000) have argued, the knowledge of appropriately rich vocabulary significantly contributes to the acquisition of various language skills, enhancing both academic writing competence (Csomay & Prades, 2018) and academic performance (Schuth *et al.*, 2017), and ultimately helping students gain acceptance into the academic discourse community (Cribb & Wang, 2021).

Lexical combinations (also known as *multiword expressions, formulaic language, phrasemes,* or *chunks*) have attracted the attention of several authors (Cortes, 2008; Neff, 2008; Römer & Arbor, 2009; Simpson-Vlach & Ellis, 2010; Wray, 2013), as their proficient use is a hallmark of expert writing and results in higher-quality academic texts. Successfully deploying lexical combinations involves mastering both lexical and syntactic elements, with phraseology serving as an interface between vocabulary and syntax. The most abundant research into lexical combinations has focused on the concept of *lexical bundles*, a term coined by Biber *et al.* (2004), e.g., 'if we look at' or 'in the case of'. This broad concept refers to empirically grounded n-grams of three to six words, with frequency as the main criterion for their identification (Salazar, 2014). Biber *et al.* (1999) defined lexical bundles as "bundles of words that show a statistical tendency to co-occur" (1999:989), regardless of their idiomaticity and structural status. In Spanish academic discourse, partially equivalent to the concept of lexical bundles is the notion of *formulae*. This term was coined by Alonso-Ramos *et al.* (2017) and García-Salido *et al.* (2018) as an adaption of the *formulemes* concept in meaning-text theory (MTT; Mel'čuk, 2012, 2015, 2020). Most research into lexical bundles has primarily focused on collecting the units most frequently used in academic texts, attributing them with discourse functions (Hyland, 2008), and including them in academic writing tools for students.

Collocations have likewise gained attention as a relevant research topic, with numerous studies in recent decades addressing the resulting learner difficulties (e.g., Nesselhauf, 2005; Durrant, 2009; Laufer & Waldman, 2011). Lists of academic collocations have resulted, e.g., the Academic Collocation List (ACL; Ackermann & Chen, 2013), the Academic English Collocation List (AECL; Lei & Liu, 2018), and the French Lexique Scientifique Transdisciplinaire (LST; Tutin, 2018). Collocation extraction and list compilation methods have been primarily grounded in the specific notion of collocation used in each case, giving rise to multiple and diverse perspectives. Lexicographers and linguists have tended to adopt one of two clearly differentiated approaches to collocation analysis, namely, statistical and phraseological.

In the statistical approach, combinations are considered significant when they co-occur more frequently than they would do by chance (Jones & Sinclair, 1974). Collocations are formed by a node (the nucleus) and a collocate, the word that typically appears within a span of three or four words to the left or right (e.g., Ackermann & Chen, 2013). The ACL and AECL both are based on the statistical approach, but applying slightly different criteria. The ACL, with a focus on collocational spans, relies on T-score and mutual information (MI)[1] association measures and on frequency and distributional filters to identify collocations in an extensive academic corpus. Similar refined association measures are used for the AECL, but greater emphasis is placed on syntactic relationships.

In the phraseological approach, a collocation is a phraseme (Mel'čuk, 2015), i.e., a non-free phrase made up of a base and a collocate. The base, semantically and intentionally selected by the speaker, imposes lexical constraints regarding the collocate. To express a particular meaning, the base thus demands a specific lexical unit, and consequently, frequency ceases to be the sole or most important criterion. For instance, in Spanish, a speaker's choice of the base *hipótesis* ('hypothesis') imposes lexical restrictions on potential collocates that means that *formular* ('formulate') is possible, but not *hacer* ('do'). Collocations are distinguished from other types of phraseological units, such as idioms, in that being compositional, so their meaning corresponds to the sum of their parts (Mel'čuk, 2012). For our CAE list we adopt a phraseological approach, specifically, the MTT framework.

[1] According to Lei and Liu (2018:227), MI measures the strength of association between two words in a collocation by considering both their co-occurrence frequency and their independent frequency, whereas the T score assesses the confidence level regarding whether an association between the two words exists. While both metrics are valuable for identifying collocations, their focus is slightly different, as MI favours low-frequency words that co-occur frequently, whereas the T score prioritizes high-frequency collocations.

Lists of academic collocations and similar resources are proposed with several practical purposes. They are included in courses to assist teachers/students in teaching/acquiring the vocabulary necessary to produce university-level texts (e.g., Lei & Liu, 2018). They are also implemented in lexicographic tools designed to help with academic writing. Examples are, for English, the Louvain English for Academic Purposes Dictionary corpus tool (LEAD; Paquot & Granger, 2012) and SciE-Lex (Laso, 2022), and for French (with a similar format), the French LST (Jacques & Tutin, 2018). Another resource is ColloCaid (Frankenberg-Garcia *et al.*, 2019; Lew *et al.*, 2018), a real-time text editor that suggests a wide variety of collocations for academic writing in English.

For the Spanish-speaking context, the corpus-dictionary tool HARTA, which includes a module with collocations, focuses explicitly on academic phraseology. We refer further to HARTA below, as it incorporates the collocations in the CAE list.

3. METHODOLOGY

3.1. Corpus

The corpus used to identify collocations used in Spanish academic discourse was HARTA-Expertos-Plus (HEP; Guzzi, 2023), compromising 21 067 836 words and 3905 research articles in two parts. The HARTA-Expertos corpus (García-Salido *et al.*, 2019), with 413 texts totalling 2 025 092 words, is based on the Spanish section of the Spanish-English Research Articles Corpus (SERAC; Pérez-Llantada, 2014) and selected scientific articles from the Red Iberoamericana de Innovación y Conocimiento Científico database. The remaining 19 042 744 words come from 3492 Corpus Iberia scientific texts (Ahumada, 2011), included in order to maximize collocation extraction from expert texts and so compile a sufficiently representative list of academic collocations in Spanish.[2]

HEP, which follows the SERAC structure, is divided into 12 disciplines in four main academic fields (Table 1) containing approximately five million words each: Arts and Humanities (AH), Biology and Health Sciences (BH), Physical Sciences and Engineering (PE), and Social Sciences and Education (SE).

Table 1. The HARTA-Expertos-Plus (HEP) academic corpus.

Field	Discipline	No. of texts	No. of words	Total words	Total texts
AH	Library Science	22	128 616	5 300 448	664
	Linguistics	255	1 986 684		
	Literature	223	1 849 091		
	Art	164	1 336 057		
BH	Biology	46	206 011	5 315 966	1448
	Health Science	1402	5 109 955		
PE	Physics & Chemistry	185	706 353	5 388 692	1052
	Earth Sciences	418	2 553 628		
	Engineering	449	2 128 711		
SE	Economy	57	375 922	5 062 730	742
	Education	282	1 855 454		
	Sociology	403	2 831 354		
Total				21 067 836	3905

To extract academic collocations, the HEP corpus was processed using a six-step methodology (Figure 1), adapted from that used for the HARTA-Expertos corpus. Texts were formatted as XML, edited to remove noise that could interfere with subsequent processing, then tokenized and lemmatized using a customized adaptation of the LinguaKit tool (Garcia & Gamallo, 2016), which detected sentence boundaries considering punctuation marks, scientific names, multiword expressions, and abbreviations. Next, part-of-speech (PoS) tagging was performed

[2] For the purposes of this study and for practical reasons, a text was classified as 'expert' if it was published in scientific journals reflecting different disciplines, i.e., author characteristics were not taken into account.

using Freeling (Padró & Stanilovsky, 2012), and finally, since this study relies on dependency syntax – specifically on Universal Dependency (UD) parsing (Nivre *et al.*, 2016) –, the UDPipe program was used to syntactically analyse the corpus (Straka *et al.*, 2016) for relationships between bases and collocates.

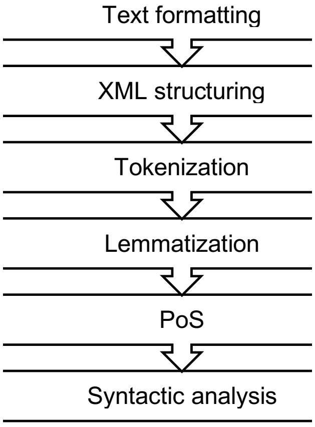

Figure 1. HEP corpus processing.

3.2. Identification of academic collocations

To identify collocations from the HEP corpus, we used refined methods as used in relevant studies, focusing particularly on a phraseological rather than a statistical approach.

3.2.1. Candidate collocation extraction

In a semi-automated process, as per previous studies (Tutin, 2018), 326 266 collocation candidates were automatically extracted from the syntactically analysed corpus, based on the five most productive syntactic dependency relations with the noun (N) as the base: N+*de*+N (nmod in UD terminology; De Marneffe *et al.*, 2021), N+Adj (amod), V+N (obj), N+V (nsubj), and N+(prep)+V (obl). In all five relations, the noun (N) is the base of the collocation. Nouns used in these combinations are drawn from a Spanish academic word-list (García-Salido, 2021). This list was identified following a corpus-driven approach from the HARTA-Expertos corpus using specificity and distribution criteria. It comprises 1239 lemmas, among which 602 are nouns.

Only collocations with a frequency threshold *f*≥5 in the HEP corpus were selected. This threshold – which resulted in more noise but also yielded more valid collocations – was established relative to the size of the HARTA-Expertos corpus (*f*≥2.2 per million words) following the approach described in the corresponding literature, and despite the larger size of the HEP corpus. Were this threshold to be adjusted (e.g., *f*≥46, or *f*≥2.2 per million words), we would have excluded valid collocations such as *corroborar hipótesis* ('corroborate hypothesis') and *ligera tendencia* ('slight trend').

Also extracted for each collocation candidate was information on seven association measures (including log-likelihood (*G2*), MI, and T score) commonly used to compile academic collocation lists. This information was not used as a primary criterion for candidate extraction, but was extracted to be able to consider the results obtained for our method in the light of insights provided by the association measures.

Once the initial 326 266 candidates were obtained, we addressed the remaining noise in the list. This involved filtering collocations (a) that complied with our phraseological criteria and (b) that were interdisciplinary in nature, i.e., excluding specialized collocations and retaining only those applicable across disciplines.

3.2.2. Manual filtering to identify collocations

A customized online tool was specifically developed to distinguish candidates corresponding to other types of lexical combinations, such as free combinations (e.g., *factor diferente,* 'different factor'), multiword units (e.g., *base de datos,* 'database'), and parts of lexical bundles (e.g., *siguiente tabla* as in *en la siguiente tabla se muestra,* 'in the following table is shown').

Collocations from each of the five syntactic groups were selected by three annotators according to our phraseological criteria. Combinations that lacked compositionality and exhibited greater fixedness than collocations were excluded (e.g., *tener lugar,* 'take place', *base de datos*, 'database', *cambio climático,* 'climatic change'). Since the distinction between free combinations and collocations is often unclear, special focus was placed on the concept of restricted lexical co-occurrence (Mel'čuk, 2015). For instance, included were collocations such as *obtener resultado* ('obtain result') and *arrojar resultado* ('yield result'), but not *modificar resultado* ('modify result'). Some other collocations, particularly those in the subj+V category (e.g., *posibilidad existe* 'the possibility increases' or *valor disminuye* 'the value decreases'), were also included, even though they were more challenging to identify due to their weaker lexical association from a statistical perspective (e.g., using MI scores).

To ensure selection reliability from a qualitative standpoint and maximum inter-annotator agreement, the selection process was divided into three phases: initial individual annotation, expert review and annotation, and a final consensus phase. Thus, two annotators made their individual selections independently and the expert annotator then reviewed those selections. The decision stood when the expert agreed as to retaining or discarding collocations, and when there was disagreement, the contentious collocations were relegated for discussion in the consensus phase, involving weekly meetings of the three annotators aimed at reaching a unified decision.

This phraseological analysis resulted in the manual selection of 5884 collocations from the 326 266 automatically extracted candidates (Table 2).

Table 2. Data collected from the filtering process (I).

	Number of collocations
After automatic extraction	326 266
After phraseological filtering	5884

3.2.3. Semi-automatic filtering to identify interdisciplinary collocations

Interdisciplinarity was the second criterion applied, given that academic collocations should be useful across disciplines and independent of specific disciplines (Drouin, 2010; Tutin, 2018), as in, e.g., *llevar a cabo un estudio* ('conduct a study') and *extraer datos* ('extract data'). Since the collocations selected in the manual filtering phase included general-language items (e.g., *jefe de departamento,* 'department head') and discipline-specific items (e.g., *ingresar paciente*, 'admit patient'), parameters were established based on frequency and distribution.

3.2.3.1. Collocation vetting

Fist, our criterion was to consider collocations as interdisciplinary if bases were not assigned to a specific discipline. Accordingly, from the initial 602 nouns in García-Salido's academic list (2021), 83 were discarded due to their uneven distribution, leaving 519 nouns. Once any specialized nouns were discarded, the remaining collocations were analysed for frequency and distribution.

To determine the optimal percentage for discarding specialized combinations several pilot tests were implemented, adjusting both distribution and standard deviation (SD) values. According to Gardner and Davies (2014:315-326), there is no universally validated value for the distribution criterion; however, we selected SD because it would indicate how well data were distributed across the four academic fields (AH, BH, PE, and SE). Thus, for a collocation to be considered interdisciplinary, it had to meet the following criteria: (a) SD between approximately 0.00 and 0.24, (b) occurrence in ≥20% of texts in at least three fields or in two fields provided one each was in PE-BH and SE-AH, and (c) occurrence in ≥3 disciplines. To counterbalance the relatively low threshold regarding disciplines (≥3 of 12), we discarded collocations occurring with high frequency (90%) in a single field, and collocations (even if well distributed) that did not occur in texts belonging exclusively to disciplines classified under the same field (e.g., health sciences and chemistry, compared to health sciences and literature).

Figure 2 compares quantitative data for examples of a genuinely interdisciplinary collocation – *arrojar dato* ('yield data') – and a specialized collocation corresponding to a very specific discipline – *recibir dosis* ('receive dose').

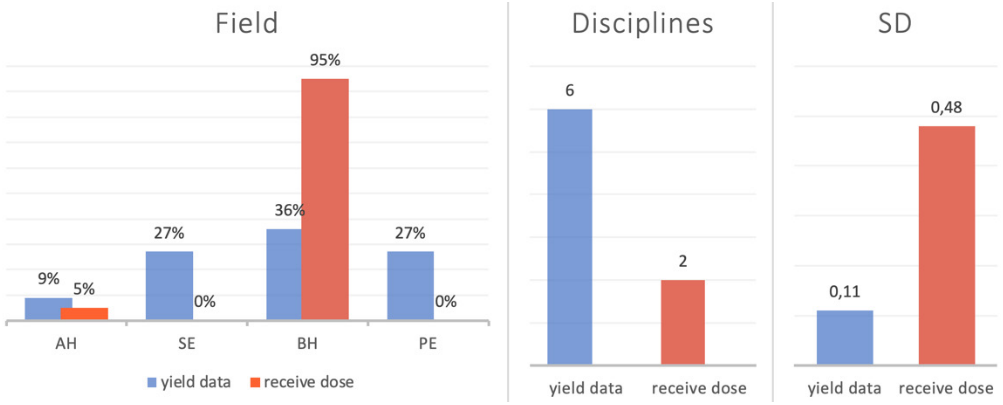

Figure 2. Data for an interdisciplinary collocation (blue) and a specialized collocation (red).

To filter academic collocations, we set the following criteria: (1) if distribution was heterogeneous, collocations with a given base were discarded, and (2) if at least 25%[3] of a specific noun-base was evenly distributed across fields, the collocation was retained, provided it did not reflect a specialized meaning that was not productive in academic discourse. Discarded, consequently, were 35 bases and their poorly distributed collocations that were considered to possibly be associated with a more prototypical meaning in specific disciplines. As a result of the interdisciplinary analysis, 118 bases and 398 collocations were discarded, leaving 5486 collocations (Table 3).

Table 3. Data collected from the filtering process (II).

	Number of collocations
After automatic extraction	326 266
After phraseological filtering	5884
After interdisciplinary filtering	5486

3.2.3.2. Polysemy filtering

Identified among the remaining collocations were polysemous bases, requiring a detailed analysis of differing meanings in line with the proposals of various authors (e.g., Frankenberg et al., 2019; Skoufaki & Petric, 2021). After an exhaustive manual analysis of meanings and contexts of the bases, for the purposes of this analysis, we selected 30 polysemous bases with only one relevant meaning in academic discourse, and 18 polysemous bases with two meanings in academic discourse.

An example of the first case is *concentración* ('concentration', referring to 'gathering in one place what was previously dispersed'). We included collocations such as *alta concentración* and *disminuir concentración* ('high concentration' and 'reduce concentration'), but, since distribution across fields was heterogeneous, we discarded *capacidad de concentración* ('ability to concentrate', which in the HEP corpus reflects only the meaning of 'the mental state that allows one to focus on a single thing'). Regarding the second case, most bases had unequivocal collocations for each meaning. An example of an ambiguous base was *medida* ('measure'), with two clearly differentiated meanings. When the meaning was 'expression of the result of a measurement', we included collocations such as *calcular medida* ('calculate measure') or *alcanzar medida* ('reach measure'), and when the meaning was 'disposition, prevention', we likewise included collocations like *medida preventiva* ('preventive measure') and *instaurar medida* ('establish measure').

As a result of this analysis, non-interdisciplinary meanings (reflecting specialized collocations corresponding to specific disciplines) were discarded, leaving 5450 collocations in the final list (Table 4).

Table 4. Data collected from the filtering process (III).

	Number of collocations
After automatic extraction	326 266
After phraseological filtering	5884
After interdisciplinary filtering	5486
After polysemy filtering	5450

[3] This threshold was established after extensive pilot testing and was deemed appropriate for the capture of collocations of interest for general academic discourse.

3.2.4. CAE list validation

A common practice on compiling specific vocabulary lists is to validate both the list and the methodology used, as done for several academic vocabulary lists proposed in the literature (e.g., Coxhead, 2000; Lei & Liu, 2018; Dang *et al.*, 2022), which use different validation options to measure list representativeness compared to other registers. To validate the CAE list, we used the Nguyen and Coxhead (2022) methodology for academic collocations as applied to the ACL and AECL, basing our analysis on the approach proposed by Lei and Liu (2018).

First, overall CAE coverage across the HEP corpus was calculated with respect to two reference corpora: the narrative part of the LEXESP corpus (Sebastián *et al.*, 2000) and the AnCora-ES general-language corpus (Taulé, *et al.*, 2008). We used the same extraction method as described above to extract collocations from the reference corpora, and validated the lists using the Nguyen and Coxhead (2022) measures (Figure 3). Second, collocations for the five most frequent bases in the CAE list (*studio,* 'study', *trabajo,* 'work', *información,* 'information', *dato,* 'data', *resultado,* 'result') were compared with the esTenTen18 Sketch Engine general-language corpus (Kilgarriff & Renau, 2013). The aim was to identify overall HEP collocation coverage in the HARTA-Expertos, LEXESP, and AnCora-ES corpora, and to determine the degree of overlap between the collocates for the five most frequent CAE bases and the collocates yielded by these corpora.

$$\text{Average coverage} = \frac{\sum \text{frequency of all list items} \times 2}{\text{Corpus size}} \times 100 \qquad \text{Average coverage} = \frac{\text{Overall coverage}}{\text{Number of list items}}$$

Figure 3. Validation measures for vocabulary lists (Nguyen & Coxhead, 2022).

Regarding the first method, the occurrence of CAE collocations in the two reference corpora was identified by extracting candidates from each corpus. Similar to Lei and Liu (2018), this analysis was restricted to the V+N, N+V, and N+Adj groups. The results obtained showed that overall academic collocation coverage was 1.52% in the academic corpus, compared to 0.10% in the narrative part of the LEXESP and 0.64% in the general-language AnCora-ES corpus. So, the CAE list covered the academic corpus 8.86 times more than AnCora-ES corpus, and 14.15 times more than the narrative part of the LEXESP corpus (Table 5).

Table 5. CAE list overlap with the HEP, LEXESP and AnCora-ES corpora.

Corpus	Tokens, n	Types, n	Overall coverage, %	Average coverage, %
HEP	161 013	4050	1.528	0.0003
LEXESP	1084	77	0.108	0.0014
AnCora-ES	1656	744	0.642	0.0008

These results align well with findings by Nguyen and Coxhead (2022), who reported that coverage in their Corpus of Contemporary American English (COCA) Fiction reference corpus was 0.06% for ACL and 0.10% for AECL compared to 0.84% for ACL and 1.46% for AECL for academic corpora. Our results are also consistent with overall coverage results for ACL and AECL, where academic collocation coverage (in our case, <1.5%) was generally low (<2.8%). As for average coverage, percentages do not reveal relevant differences, as percentages in the HEP and AnCora-ES corpora were similar (0.0003% and 0.0008%, respectively), and slightly higher in the LEXESP corpus (0.0014%) though not as much when compared to overall coverage. Moreover, as suggested by Nguyen and Coxhead (2022:103), findings on this measure can be problematic given the frequency of words. However, note that our average coverage index is in line with the index reflecting ACL and AECL in the COCA Academic corpus (around 0.0003%), suggesting that academic collocations cover a similar portion of texts regardless of language.

Regarding the second method, we identified esTenTen18 collocates using the Word Sketch tool, and – corroborating Lei and Liu's (2018) results – we found that the collocates that overlapped were not exact equivalents for any of the five bases. For instance, for collocations with *resultado,* in HEP we obtain *constatar resultado* ('ascertain result') and *resultado nulo* ('null result'), while in esTenTen18 we obtain *resultado impresionante* ('impressive result') and *saber resultados* ('know results'), i.e., collocations are not shared by both corpora. While Lei and Liu (2018) reported that 70% of collocates for selected AECL bases are not found in general-English collocation dictionaries, we obtained a lower percentage, ranging from 18% to 36%. Despite the lower percentage, a significant portion of collocates that appear to be more exclusive to academic discourse were represented. Additionally, the percentage rises to around 80% when we consider collocates that appear in the general-language corpus but were not selected in HEP. This may be due to the method used to select collocations (phraseological versus statistical), but more so perhaps, due to a greater restriction on collocations in academic discourse given the specificity of their use in this register.

4. RESULTS

The CAE list currently consists of 5450 collocations implemented in the HARTA lexical tool (Guzzi & Alonso-Ramos, 2023a), a number approximately midway between totals for the LST, ACL, and AECL, which have 1600, 2468, and 9029 collocations, respectively. The explanation for the differences is that different methods were used to identify collocations. An excerpt from the CAE list is shown in Table 6.

Table 6. CAE listing for the noun posibilidad ('possibility').

	Syntactic relationship	Collocations
posibilidad	~ + Adj	*posibilidad escasa*
		posibilidad infinita
		posibilidad mayor
	N de ~	*conjunto de posibilidad*
	~ + V	*posibilidad aumentar*
		posibilidad existir
	V + ~	*abrir posibilidad*
		aceptar posibilidad
		agotar posibilidad
		contemplar posibilidad
		multiplicar posibilidad
		ofrecer posibilidad
		plantear posibilidad
		presentar posibilidad
		proponer posibilidad
		proporcionar posibilidad
		tener posibilidad
	V + prep ~	*apuntar (a) posibilidad*
		contar (con) posibilidad

The number of collocations per syntactic group is summarized in Table 7. The V+N group contains the largest number of collocations, followed by the N+Adj groups (around a third each), reported in the literature to be the two syntactic groups that pose the greatest challenge for L2 learners (Nesselhauf, 2005; Laufer & Waldman, 2011; Ackerman & Chen, 2013; Lei & Liu, 2018). Almost 70% of the CAE list thus primarily contains collocations reflecting the two most challenging groups for learners and users of academic Spanish.

Table 7. Quantitative breakdown of the CAE list.

	N	%	Collocation example
N+Adj	1778	33%	*objetivo principal* ('main goal')
N+*de*+N	836	15%	*objeto de investigación* ('object of research')
V+N	1886	35%	*proporcionar ejemplo* ('provide example')
V+(prep)+N	564	10%	*llegar a resultado* ('obtain result')
N+V	386	7%	*autor aboga* ('author advocates')
TOTAL	5450		

The N+Adj group (33% of the CAE) includes collocations that help students precisely evaluate and describe academic nouns. As Tutin (2014) pointed out, some adjectives in N+Adj collocations express an evaluation, e.g., *estudio principal* ('main study'), *artículo excelente* ('excellent article'), or *resultado significativo* ('significant result'). Others reflect a more objective and descriptive character of the nouns they modify, e.g., expressing size/quantity (*gran volume,* ‘large volume’ and *alta frecuencia*, ‘high frequency’), or signalling specificity, given that concepts in academic discourse tend to be expressed concretely (*estudio cuantitativo*, ‘quantitative study’ and *estudio descriptivo*, ‘descriptive study’).

Common within the group of V+N collocations (35% of the CAE) are support verb constructions (e.g., *hacer un studio,* ‘conduct a study’, *llegar a una conclusion*, ‘reach a conclusion’), due to the high frequency of nominalizations in Spanish academic discourse. Other constructions include collocations with full lexical verbs (e.g., *consultar studio*, ‘check a study’, *evaluar la competencia*, ‘evaluate competence’), collocations formed by aspectual verbs and causative verbs (e.g., *cerrar apartado*, ‘close section’*, acometer estudio*, ‘undertake study’), which contextualize the phase of the noun (beginning/end). The list also includes collocations with causative values, such as *generar condición* (‘generate condition’) and *elaborar cuestionario* (‘draft questionnaire’).

The N+*de*+N and the V+(prep)+N collocations represent 15% and 10%, respectively, of the total. In the N+*de*+N group, the base is the second noun, and illustrated is a semantic ‘part/unit of’ or ‘level of’ relationship, as in e.g., *miembro de equipo* (‘member of team’), *grado de aceptación* (‘degree of acceptance’), and *fase de análisis* (‘phase of analysis’). In fact, very frequent in academic texts are collocates that add a ‘level of’ meaning to the base, inherently expressing gradation in their lexicographic definition. As for the V+prep+N group, the noun converges with two types of complements at the syntactic-semantic level: arguments, i.e., verb complements governed by prepositions, and obliques, which are introduced by verbs that do not obligatorily govern a preposition. Although nouns functioning as governed complements resemble direct objects at the syntactic level, all types of collocations that link both elements through a preposition were included in this group in the CAE list. Selected were collocations that already appear in the V+N group as objects but with a syntactic shift that does not alter meaning, e.g., *presentar* (*como*) *alternativa* (‘present (as) an alternative’) or *establecer* (*como*) *objetivo* (‘establish (as) an objective’).

Finally, the group containing the fewest collocations (7%) was the N+V group, where the noun functions as the grammatical subject. Most such collocations occur in examples like *autor defiende* (‘author supports’), representing prototypical actions of the noun as subject. Although the number of collocations in this group is relatively small (n=386), these combinations can help students acquire useful expressions for academic writing, particularly those referring to reporting verbs (e.g., *autor apunta, sugiere, aboga*, ‘author points out, suggests, advocates’*,* etc). Furthermore, as noted by Lei and Liu (2018), a significant additional contribution of this collocational group is that it demonstrates that certain nouns in academic discourse can be used metonymically, i.e., an inanimate object (e.g., *article*) can be treated as an animate object and so can function as the subject of actions that are more prototypically performed by people, e.g., *artículo analiza* (‘article analyses’) and *bibliografía sugiere* (‘literature suggests’).

5. DISCUSSION

Addressing the validity of the results obtained using our extraction method, data on association measures were considered in order to compare our results with those of other methods typically employed in collocation extraction processes. In order to compare both methods and assess the degree of overlap between the selected 5884 Spanish academic collocations, we briefly analysed *G2* and MI association measure information extracted for the collocations selected in the initial phase.

For this analysis, following Tutin (2018), we used *G2* for a threshold of ≥10.7 (p<0.001), and MI for a threshold of >3, as used for the AECL and ACL collocations. We found that 37% of our selected collocations would have been discarded if *G2* was employed with a threshold of p<0.001, and an even more substantial 48% if MI was employed (2180 and 2836, respectively, of the total). Jointly, the percentage of collocations discarded was 30%, considering that 95% of *G2* discards coincide with MI discards. Thus, although MI is more restrictive, it seems to broadly align with *G2* in terms of discarded collocations.

Considering the collocations selected by both measures, the degree of overlap between the selected collocations was low and so use of both measures would result in the loss of many valid collocations frequent in academic discourse. To evaluate the results from a qualitative perspective, we analysed examples of the collocations comprising each case. Table 8 illustrates two examples each, as follows: included on the basis of three criteria (Case 1), included on the basis of two criteria (Cases 2 and 3), included according to phraseological criteria (Case 4), and excluded on the basis of three criteria (Case 5).

Table 8. Examples of collocations excluded and included following different approaches.

Case	Collocations	Phraseological criteria	MI	*G2*
1	*desempeñar una función* ('perform a function'); *vital importancia* ('vital importance')	in	in	in
2	*argumento fundamental* ('fundamental argument'); *trabajo ingente* ('enormous effort')	in	in	out
3	*estimar el impacto* ('estimate impact'); *aplicación de una metodología* ('application of methodology')	in	out	in
4	*ahondar en un estudio* ('delve into a study'); *medir parámetro* ('measure a parameter')	in	out	out
5	*normalizar una función* ('normalize a function'); *importancia normal* ('normal importance')	out	out	out

Nonetheless, considering the classification and discard percentages for each association measure, we believe that, for Spanish academic collocations, *G2* with a slightly lower threshold (p<0.01) could produce results similar to those obtained according to phraseological criteria, given that the collocations with higher *G2* indices (>2000) included high-frequency collocations – e.g., *largo plazo* ('long term'), *factor de riesgo* ('risk factor'), *toma de decisión* ('decision-making'), *formar parte* ('be part of') – that largely corresponded to the collocations selected for the CAE list.

6. CONCLUSIONS

With the aim of contributing to research into Spanish for Academic Purposes, we developed a reference list of academic collocations in Spanish (the CAE list), describing in detail the methodology and criteria followed to identify the collocations.

In terms of its applications, the CAE list can play a crucial role in setting vocabulary goals at university level. For students writing academic Spanish, it can be used as a reference tool to consult whenever uncertainty exists regarding the appropriateness of frequent collocations used in this area. It can also be used as a reference resource for the development of teaching and learning materials and for online practice materials.

Regarding online lexical resources, the full CAE list has been implemented in HARTA, along with quantitative information on collocation use, based on our interdisciplinary analysis, in the different scientific disciplines covered by the HEP corpus. This information is useful for students to gain insights into whether a collocation is more typical of a field, even if is still considered a general-academic collocation, and for researchers interested in the study of academic collocations. Moreover, the analysis of academic meanings has led to the inclusion of a new functionality in the collocations panel, allowing access to multiple meanings of the base and to the corresponding collocations. It is important to emphasize that users should both learn and use the different academic meanings of lexical elements.

The CAE list can be also used to develop assessment materials to measure academic collocation use and knowledge of academic vocabulary. The inclusion of the CAE list of collocations in vocabulary tests is supported by research demonstrating that knowledge of collocations is the most significant indicator of vocabulary competence (Crossley *et al.*, 2015). Indeed, an analysis of academic collocational complexity has been addressed by Guzzi & Alonso-Ramos (2023b) and Guzzi (2023), along with a beta proposal for a collocation complexity tool for academic Spanish.

Regarding implications for future research, the CAE list has been designed for use by both native Spanish speakers and L2 learners. However, since L2 difficulty in learning collocations is likely to vary depending on the learner's L1, it may be useful to develop academic collocation lists for learners from specific L1 groups. Further improvements that would potentially improve students' use of academic vocabulary could be to analyse which collocations are frequently underused or are used incorrectly, and to collect and analyse other types of university texts to identify lexical gaps. The results could be included in a database of collocational lapses and typical uses that could benefit HARTA users, possibly categorized and adapted to different L1s.

Finally, although the CAE list is comprehensive in including 5450 collocations derived from a corpus exceeding 21 million words, it should be regarded as a dynamic compilation to be updated with future findings of research into Spanish academic collocations. This includes potential additions or exclusions prompted by insights gained from broader academic corpora.

REFERENCES

Ackermann, K., & Chen, Y. (2013). “Developing the Academic Collocation List (ACL) –A corpus-driven and expert-judged approach”, *Journal of English for Academic purposes* 12/4, 235-247. https://doi.org/10.1016/j.jeap.2013.08.002

Ahumada, I. (2011). “El español de la ciencia: ¿la identidad en crisis”, in *Word for Word/Palabra por palabra. El impacto social, económico y político del español y del inglés.* Madrid: Santillana Español-British Council-Instituto Cervantes, 309-328.

Alonso-Ramos, M., García-Salido, M., & Garcia, M. (2017). “Exploiting a corpus to compile a lexical resource for academic writing: Spanish lexical combinations”, in I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček & V. Baisa (eds.) *Electronic lexicography in the 21st century, Proceedings of 2017 eLex Conference*, 571-586. https://elex.link/elex2017/wp-content/uploads/2017/09/paper35.pdf

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999*). Longman grammar of spoken and written English*. Harlow: Pearson Education Limited.

Biber, D., Conrad, S., & Cortes, V. (2004). “If you look at…: Lexical bundles in university teaching and textbooks”, *Applied linguistics* 25/3, 371-405. https://doi.org/10.1093/applin/25.3.371

Boers, F., & Webb, S. (2018). “Teaching and learning collocation in adult second and foreign language learning”, *Language Teaching* 51/1, 77-89. https://doi.org/10.1017/S0261444817000301

Cortes, V. (2008). “A comparative analysis of lexical bundles in academic history writing in English and Spanish”, *Corpora* 3/1, 43-57. https://doi.org/10.3366/E1749503208000063

Coxhead, A. (2000). “A new academic word list”, *TESOL Quarterly* 34/2, 213-238. https://doi.org/10.2307/3587951

Cribb, M., & Wang, X. (2021). “Making academic vocabulary count through strategic deployment in oral presentations by Chinese students of English”, *The Language Learning Journal* 49/2, 251-264. https://doi.org/10.1080/09571736.2019.1566396

Csomay, E., & Prades, A. (2018). “Academic vocabulary in ESL student papers: A corpus-based study”, *Journal of English for Academic Purposes* 33, 100-118. https://doi.org/10.1016/j.jeap.2018.02.003

Crossley, S.A., Salsbury, T., & McNamara, D. (2015). “Assessing lexical proficiency using analytic ratings: A case for collocation accuracy”, *Applied Linguistics* 36/5, 570-590. https://doi.org/10.1093/applin/amt056

Da Cunha, I., Montané, M.A., & Hysa, L. (2017). “The arText prototype: An automatic system for writing specialized texts”, in A. Martins & A. Peñas (eds.) *EACL 2017 15th Conference of the European Chapter of the Association for Computatinal Linguistics. Proceedings of the Software Demonstrations.* Valencia: ACL (Association for Computational Linguistics), 57-60. http://hdl.handle.net/10230/46442

Dang, T.N.Y., Webb, S., & Coxhead, A. (2022). “Evaluating lists of high-frequency words: Teachers’ and learners’ perspectives”, *Language Teaching Research* 26/4, 617-641. https://doi.org/10.1177/136216882091118

De Marneffe, M., Manning, C.D., Nivre, J., & Zeman, D. (2021). “Universal Dependencies”, *Computational Linguistics* 47/2, 255–308. https://doi.org/10.1162/coli_a_00402

Drouin, P. (2010). “Extracting a bilingual transdisciplinary scientific lexicon”, in S. Granger & M. Paquot (eds.) *eLexicography in the 21st century: new challenges, new applications.* Louvain-la-Neuve: Presses Universitaires de Louvain/Cahiers du CENTAL, 43-53.

Durrant, P. (2009). “Investigating the viability of a collocation list for students of English for academic purposes”, *English for Specific Purposes* 28/3, 157-169. https://doi.org/10.1016/j.esp.2009.02.002

Durrant, P. (2016). “To what extent is the Academic Vocabulary List relevant to university student writing?”, *English for specific purposes* 43, 49-61. https://doi.org/10.1016/j.esp.2016.01.004

Errázuriz Cruz, M.C. (2014). “El desarrollo de la escritura argumentativa académica: los marcadores discursivos”, *Onomázein* 30, 217-326. https://doi.org/10.7764/onomazein.30.13

Frankenberg-Garcia, A. (2018). “Investigating the collocations available to EAP writers”, *Journal of English for Academic Purposes* 35, 93-104. https://doi.org/10.1016/j.jeap.2018.07.003

Frankenberg-Garcia, A., Lew, R., Roberts, J.C., Rees, G.P., & Sharma, N. (2019). “Developing a writing assistant to help EAP writers with collocations in real time”, *ReCALL* 31/1, 23-39. https://doi.org/10.1017/S0958344018000150

García-Salido, M., García-González, M., & Alonso-Ramos, M. (2019). “Identifying lexical bundles for an academic writing assistant in Spanish”, in G. Corpas Pastor & R. Mitkov (eds.). *Computational and Corpus-Based Phraseology, Third International Conference, Europhras 2019*. Málaga: Springer, 144-158. https://doi.org/10.1007/978-3-030-30135-4_11

García-Salido, M., Garcia, M., Villayandre-Llamazares, M., & Alonso-Ramos, M. (2018). “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)*. Paris: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/pdf/769.pdf

García-Salido, M. (2021). “Compiling an Academic Vocabulary List of Spanish”. Retrieved from: https://doi.org/10.13140/RG.2.2.27681.33123

Garcia, M., & Gamallo, P. (2016). “Yet another suite of multilingual NLP tools”, in J.P. Leal, J.L. Sierra-Rodríguez & A. Simões (eds.) *Languages, Applications and Technologies. Communications in Computer and Information Science*. Madrid: Springer, 65–75. https://gramatica.usc.es/~gamallo/artigos-web/SLATE2015.pdf

Gardner, D., & Davies, M. (2014). “A new academic vocabulary list”, *Applied linguistics* 35/3, 305-327. https://doi.org/10.2307/3587951

Guzzi, E. (2023). *Identificación automática de colocaciones académicas en español para una herramienta en línea de ayuda a la redacción.* Tesis doctoral. http://hdl.handle.net/2183/35240

Guzzi, E., & Alonso-Ramos, M. (2023a). “Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español”, *Tecnologías para la investigación en segundas lenguas 2*, 1-22. https://doi.org/10.1344/teisel.v2.42173

Guzzi, E., & Alonso-Ramos, M. (2023b). “Sofisticación y diversidad como medidas de complejidad léxica para determinar el perfil colocacional de textos académicos en español”, *Revista Signos 56*/112, 282-305. https://doi.org/10.4067/S0718-09342023000200282

Hu, M., & Nation, P. (2000). “Unknown vocabulary density and reading comprehension”, *Reading in a Foreign Language* 13/1, 403-430.

Hyland, K. (2006). “The ‘other’ English: Thoughts on EAP and academic writing”, *The European English Messenger* 15/2, 34-38. https://www.academia.edu/40422292/The_other_English_thoughts_on_EAP_and_academic_writing

Hyland, K. (2008). “Metadiscourse: Mapping interactions in academic writing”, *Nordic Journal of English Studies* 9/2, 125-143. https://doi.org/10.35360/njes.220

Hyland, K., & Tse, P. (2007). “Is there an “Academic Vocabulary”?”, *TESOL Quarterly* 412/2, 235-253.

Jacques, M.P., & Tutin, A. (2018). *Lexique transversal et formules discursives des sciences humaines*. London: ISTE Group.

Jones, S., & Sinclair, J. (1974). “English lexical collocations. A study in computational linguistics”, *Cahiers de lexicologie* 24/25, 15–61.

Kilgarriff, A., & Renau, I. (2013). “esTenTen, a vast web corpus of Peninsular and American Spanish”, *Procedia-Social and Behavioral Sciences* 95, 12-19. https://doi.org/10.1016/j.sbspro.2013.10.617

Laso, N.J. (2022). “SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English”, *TEISEL. Tecnologías para la investigación en segundas lenguas* 1*,* 1-16. https://doi.org/10.1344/teisel.v1.37444

Laufer, B., & Waldman, T. (2011). “Verb-noun Collocations in Second Language Writing: A Corpus Analysis of Learners’ English”, *Language Learning* 67/2, 647-672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Lei, L., & Liu, D. (2018). “The academic English collocation list: A corpus-driven study”, *International Journal of Corpus Linguistics* 23/2, 216-243. https://doi.org/10.1075/ijcl.16135.lei

Lew, R., Frankenberg-Garcia, A., Rees, G.P., Roberts, J.C., & Sharma, N. (2018). “ColloCaid: A real-time tool to help academic writers with English collocations”, in J. Cibej, V. Gorjan, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress*. Ljubljana: Ljubljana University Press, Faculty of Arts, 167-168.

Mauranen, A., Hynninen, N., & Ranta, E. (2016). “English as the academic lingua franca”, in K. Hyland & P. Shaw (eds.) *The Routledge handbook of English for academic purposes*. London/New York. Routledge, 44-55. https://doi.org/10.4324/9781315657455

Mel’čuk, I. (2012). *Semantics: From Meaning to Text*. [Vol. 1.] Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1515/phras-2012-0003

Mel’čuk, I. (2015). “Clichés, an understudied subclass of phrasemes”, *Yearbook of Phraseology* 6/1 55-86. https://doi.org/10.1515/phras-2015-0005

Mel’čuk, I. (2020). “Clichés and pragmatemes”, *Neophilologica* 32, 9-20. 10.31261/NEO.2020.32.01.

Nazar, R., & Renau, I. (2023). “Estilector: un sistema de evaluación automática de la escritura académica en castellano”, *Perspectiva Educacional*, 62/2, 37-59. https://doi.org/10.4151/07189729-vol.62-iss.2-art.1427

Neff, J. (2008). “Contrasting English-Spanish interpersonal discourse phrases: A corpus study”, in *Phraseology in foreign language learning and teaching*. Amsterdam/Philadelphia: John Benjamins, 85-99.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam/Philadelphia: John Benjamins.

Nguyen, T.M.H., & Coxhead, A. (2022). “Evaluating multiword unit word lists for academic purposes”, *ITL-International Journal of Applied Linguistics* 174/1, 83-111. https://doi.org/10.1075/itl.21041.ngu

Nivre, J., De Marneffe, M.C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., Mcdonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). “Universal dependencies v1: A multilingual treebank collection”, in N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Tenth International Conference on Language*

Resources and Evaluation (LREC'16). Portoro: European Language Resources Association. 1659-1666. https://aclanthology.org/L16-1262

Padró, L., & Stanilovsky, E. (2012). "Freeling 3.0: Towards wider multilinguality", in N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & St. Piperidis (eds.) *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Paris: European Language Resources Association (ELRA). 2473–2479. http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf

Paquot, M. (2012). "The LEAD dictionary-cum-writing aid: an integrated dictionary and corpus tool", in S. Granger & M. Paquot (eds.) *Electronic lexicography.* Oxford: Oxford University Press, 136-186. 10.1093/acprof:oso/9780199654864.003.0009.

Paquot, M., & Granger, S. (2012). "Formulaic language in learner corpora", *Annual Review of Applied Linguistics* 32, 130-149. https://doi.org/10.1017/S0267190512000098

Pérez-Llantada, C. (2014). "Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage", *Journal of English for Academic Purposes* 14, 84-94. https://doi.org/10.1016/j.jeap.2014.01.002

Römer, U., & Arbor, A. (2009). "English in academia: Does nativeness matter", *Anglistik: International Journal of English Studies* 20/2, 89-100. https://lexically.net/wordsmith/corpus_linguistics_links/Anglistik_2009_nativeness_89_R%C3%B6mer.pdf

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing*. Oxford: University of Oxford.

Schuth, E., Köhne, J., & Weinert, S. (2017). "The influence of academic vocabulary knowledge on school performance", *Learning and Instruction* 49, 157-165. https://doi.org/10.1016/j.learninstruc.2017.01.005

Sebastián, N., Carreiras, M.F., Cuetos, F., & Martí, M.A. (2000). *LEXESP: Léxico informatizado del español*. Barcelona: Universitat de Barcelona.

Simpson-Vlach, R., & Ellis, N.C. (2010). "An academic formulas list: New methods in phraseology research", *Applied linguistics* 31/4, 487-512. https://doi.org/10.1093/applin/amp058

Skoufaki, S., & Petrić, B. (2021). "Exploring polysemy in the Academic Vocabulary List: A lexicographic approach", *Journal of English for Academic Purposes* 54/101038. https://doi.org/10.1016/j.jeap.2021.101038

Straka, M., Hajic, J., & Strakov, J. (2016). "UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing", in N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J Odijk & S. Piperidis (eds.) *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16).* Portoro: European Language Resources Association, 4290-4297. http://www.lrec-conf.org/proceedings/lrec2016/pdf/873_Paper.pdf

Taulé, M., Martí, M.A., & Recasens, M. (2008). "AnCora: Multilevel annotated corpora for Catalan and Spanish", in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (eds.) *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).* Paris: European Language Resources Association (ELRA), 96-101. http://www.lrec-conf.org/proceedings/lrec2008/pdf/35_paper.pdf

Townsend, D., Filippini, A., Collins, P., & Biancarosa, G. (2012). "Evidence for the importance of academic word knowledge for the academic achievement of diverse middle school students", *The Elementary School Journal* 112/3, 497-518. https://doi.org/10.1086/663301

Tutin, A. (2014). «La phraséologie transdisciplinaire des écrits scientifiques: des collocations aux routines sémantico-rhétoriques», in A. Tutin y F. Grossmann (eds.) *L'écrit scientifique: du lexique au discours. Autour de Scientext*. Rennes: PUR, 27-44.

Tutin, A. (2018). «Les expressions polylexicales transdisciplinaires dans les articles de recherche en sciences humaines: retour d'expérience», in M.P. Jacques & A. Tutin (eds.) *Lexique transversal et formules discursives des sciences humaines*. London: ISTE Group, 73-90.

Wray, A. (2013). "Formulaic language", *Language teaching* 46/3, 316-334. https://doi.org/10.1017/S0261444813000013

Yao, G. (2022). *Metadiscourse use in Spanish academic writing: exploring the interface of nativeness and expertise.* Tesis doctoral. Murcia: Universidad de Murcia. http://hdl.handle.net/10201/117507