Compiling an Academic Vocabulary List of Spanish

Preprint · June 2021

DOI: 10.13140/RG.2.2.27681.33123

Marcos García Salido
University of A Coruña
41 PUBLICATIONS 107 CITATIONS

Compiling an Academic Vocabulary List of Spanish

Marcos García Salido

This paper focuses on the process of compiling an academic vocabulary list for Spanish that will provide the content for an academic writing-aid. Two basic approaches for automatic extraction of academic vocabulary appear in the literature: one excludes frequent general vocabulary from vocabulary units that are also frequent and evenly distributed in academic texts, while the other tries to identify units that occur significantly more often in academic texts than in non-academic ones. We applied these two methods in creating two candidate lists of Spanish academic vocabulary and compared them in two respects: their coverage in academic and non-academic texts and the dispersion of their items across different academic domains. The paper also addresses the issue of whether the vocabulary included in such lists is equally distributed in those domains. Finally, we propose a list that represents a compromise between good coverage of academic texts and academic specificity.

1. INTRODUCTION

Academic vocabulary lists are lexical repertoires that include vocabulary specific of academic texts—either absent from the more frequent vocabulary of a given language or particularly productive in academic discourse. Additionally, the compilers of such lists normally intend for them to be cross-disciplinary, that is, they should include vocabulary characteristic of academic texts while excluding terms of specific disciplines (Coxhead, 2000; Drouin, 2007; Gardner and Davies, 2004). The first lists of academic vocabulary were compiled mostly with a view to planning courses and developing teaching materials (Coxhead, 2000). More recently, they have been also used in the creation of resources helping academic written production (Frankenberg et al., 2018; Granger and Paquot, 2015; Tutin and Jacques, 2018; Verdaguer et al., 2013).

To the best of our knowledge, no academic vocabulary list of Spanish has yet been produced, even though academic discourse has attracted an increasing interest from linguistic research in Spanish in recent years (see Montolío, 2014, and references therein). This paper focuses on the process of compiling a list of academic vocabulary for this language and compares the validity of lists resulting from different approaches with a particular aim in mind: providing the vocabulary necessary for an academic writing assistant (see García Salido *et al.*, 2018). This tool is envisaged as providing users with suggestions on lexical combinations that are common in academic writing. In this sense, the list of academic vocabulary is just a starting point, and subsequently will be complemented with combinatorial information.

The paper is organized as follows: section 2 reviews research on the elaboration of vocabulary lists for different languages. Section 3 presents the methodology used in this investigation. Section 4 gives an account of our results and discusses them in terms of their coverage of different academic fields. The last section presents some conclusions and implications regarding the use of academic vocabulary lists.

2. ACADEMIC VOCABULARY LISTS: RESEARCH BACKGROUND

In spite of the wealth of corpus-based lexical repertoires of Spanish (Almela et al., 2005; Buchanan, 1927; Cuetos, 1995; Davies, 2006; García Hoz, 1953; Juilland and Chang-Rodríguez, 1964; Keniston, 1920; Rodríguez Bou, 1952; Sebastián-Gallés et al., 2000), the focus of these resources is placed on general rather than academic vocabulary. It is therefore interesting to review the cases of languages such as English or French to get a picture of the different approaches adopted in the development of academic lists, given the amount of attention paid to developing such resources for these languages.

In the case of English, the first academic vocabulary lists appeared in the 1970s (Coxhead, 2000: 214). In the 1980s, Xue and Nation (1984) published a University Word List, resulting from the merger of four previous lists. One of the most influential academic lists to the date is Coxhead's (2000), which originated as an updated alternative to the previous lists. This Academic Word List (AWL) is based on a 3.5 million-word corpus divided into four sections: Arts, Commerce, Law and Science. From this corpus, Coxhead extracted words with a frequency of at least a hundred (i.e., ca. 30 times per million words) that occurred at least 10 times in the four sections of the corpus. Additionally, these words had to be different from the 2k most frequent families of the West's General Service List (GSL). This process yielded a list of 570 entries, covering between 9.1% (Arts) and 12% (Commerce) of texts. When combined with West's GSL, however, the AWL coverage increased, but also the differences between the coverage of different domains—with the Science subcorpus left behind. This is one of the reasons that led Hyland and Tse (2007) to voice their reservations about all-encompassing academic vocabulary lists. Partly in response to Hyland and Tse's criticism, Coxhead and Hirsch (2007) compiled a new list extracted from a Science corpus. The method here was also subtractive: potential candidates had to be absent from GSL and AWL; additionally, they had to reach frequency and dispersion thresholds.

Despite Hyland and Tse's reservations about cross-disciplinary academic lists, further lexical repertoires of academic English have been compiled after Coxhead's AWL: Paquot's Academic Keyword List (AKL; Paquot, 2010), Gardner and Davies' New Academic Vocabulary List (NAVL; Gardner and Davies, 2014) and the lemmas of the *Oxford Learner's Dictionary of Academic English* (Lea, 2014) are cases in point. Paquot used a corpus composed of texts written by professionals and students. The “professional” subcorpus was divided into two large sections (“hard sciences” and “soft sciences”) comprising approximately two million words (800,000 and 1.2 million respectively). To this one must add ca. one million words of student writing, including texts from the humanities and social sciences along with discursive essays. In addition to controlling for the specificity of the list by including only those words that were significantly more frequent (log-likelihood ≤0.01) in the academic corpus than in the contrasting corpus (fiction texts), the words included in the list had to reach a minimum range and their distribution had to be even (Juilliand’s D ≥ 0.8). A number of words that did not meet the evenness threshold were added, based on their semantic closeness to the words already included on the list. The result of this whole process was a list of some 930 items, spanning the word classes noun, verb, adjective, adverb, and a fifth class made up mostly of pragmatic and discourse markers.

Gardner and Davies (2014) present the size of their corpus (120 million words) as an improvement with respect to previous attempts. In creating the New Academic Vocabulary List (NAVL) they established the keyness of its items by means of frequency: academic words had to be at least 50% more frequent in the academic corpus than in the contrast corpus (non-academic texts of the COCA corpus). The academic words also had to meet several criteria related to evenness of distribution: they had to reach at least 20% of their expected frequency in seven out of the nine subsections of the corpus and they also had to attain a Juilliand’s D value of 0.8 or higher. Finally, to avoid technical words, items of the list could not surpass three times their expected frequency in any particular corpus subdivision. The resulting list contains 3,000 lemmas. In order to give an idea of the coverage of NAVL, the authors subsequently grouped the lemmas into word families and created two random lists of 570 families each from the highest ranked families—the same number of word families as those contained in Coxhead’s AWL. The resulting lists covered between the 19% and 23% of the academic texts of the corpus, almost doubling Coxhead’s results.

French academic vocabulary lists have also been available for quite a while. Phal’s (1971) list is based on a corpus of nearly 2 million words composed of secondary school handbooks. Drouin (2007) compiled an academic keyword list for French based on a 2.3 million word corpus of doctoral dissertations. Like other lists previously cited, Drouin controlled for the specificity and dispersion of the list’s items. For the former he used Lafon’s (1980) calculation of specificity and drew on newspaper texts as a contrast corpus. To control for dispersion, he established a range threshold of 50 out of the 100 fragments into which his corpus was divided (the divisions were arbitrary, i.e. they had nothing to do with scientific fields). This procedure resulted in an academic list of 1,113 academic lemmas. After Drouin’s list, the first version of a cross-disciplinary vocabulary of French appeared, compiled as part of the Scientext project (Tutin, 2013). This list was based on a 2 million word corpus comprising reports, theses and articles on linguistics, economics and medicine. The criterion used to identify academic vocabulary is simpler than in other cases: a frequency threshold in all three fields. The Scientext research project has more recently produced further lexical resources restricted to the Humanities and Social Sciences domains (Jacques and Tutin 2018).

There are several projects that have compiled—or are in the process of doing so—academic vocabulary lists for other European languages. In the case of Portuguese, an academic list has been compiled manually, taking Coxhead's (2000) AWL as a guide in terms of the meanings that should be included (Baptista et al., 2010). Johansson Kokkinakis et al. (2012) have proposed different methods for compiling lists for Danish, Norwegian and Swedish, ranging from the translation of existing lists to corpus-based procedures.

Although academic lists have been designed primarily with teaching purposes in mind (Coxhead, 2000; Gardner and Davies, 2014) and have been exploited for assessment tasks (Csomay and Prades, 2018; Goodfellow et al., 2002; Laufer and Nation, 1995), they are valuable resources for Lexicography too. Paquot (2010: 206ff) already emphasised the role of these lists in the design of dictionaries and other electronic resources, such as writing aids. In fact, the AKL devised by Paquot herself has served as the basis for the macrostructure of the *Louvain English for Academic Purposes Dictionary* (LEAD; Granger and Paquot, 2015: 123). Coxhead's (2000) AWL has been used as a sort of filter to choose the lexical combinations included in the Scie-Lex dictionary (Verdaguer et al. 2013). ColloCaid (Frankenberg et al., 2018) is a writing aid under development that aims at offering its users collocational suggestions for academic English. It also uses academic vocabulary lists to identify productive collocation nodes in academic discourse, namely a combination of three lists: the core of Gardner and Davies' NAVL identified by Durrant (2016), Paquot's AKL, and the headwords of the Academic Collocation List (Ackerman and Chen, 2013). Frankenberg et al. (2018) final list includes 513 lemmas, including nouns, verbs and adjectives.

From the above reviewed studies tho different approaches emerge. We will call the former approach *subtractive,* which is followed in Coxhead (2000): she subtracts from the candidates to academic word families those that are amongst the most frequent 2k families of the GSL. In contrast, more recent proposals (Drouin, 2007; Gardner and Davies, 2014; Paquot, 2010) prefer to maintain common lemmas/word families as long as they are characteristic of academic discourse, in what we will call the *keyword* approach. This has consequences for the possible applications of academic lists. Coxhead's AWL was conceived of as a complement to the GSL. This complement can be an economic solution for the comprehension of academic texts, since it is relatively short and concentrates word families that could not seem very productive judging after frequency rankings of general English. However, the effectiveness of the AWL depends on a prior command of the most frequent word families of the GSL. Likewise, this lack of autonomy could lead to *lacunae* if the AWL were used alone, particularly for writing-related needs. Last, it disregards the fact that, in academic texts, a given lemma or word family can express senses that it rarely conveys in non-academic discourse (Gilquin et al., 2007: 324).

As for their common aspects, all of the reviewed proposals try to exclude discipline-specific vocabulary by controlling for the even distribution of the lists' items. This means that, save exceptions like Coxhead and Hirsch (2007), all the reviewed lists have a cross-disciplinary scope. For Hyland and Tse (2007) this cross-disciplinary scope is one of the most problematic aspects of academic vocabulary lists. The unequal distribution of academic vocabulary across texts from different disciplines can be considered a drawback for the cross-disciplinary validity of these repertoires. Hyland and Tse (2007) tested Coxhead's AWL in a corpus divided into three fields (Engineering, Sciences and Social Sciences), and found that only about a third of the most frequent word families were comparably frequent in all three of the domains. The least frequent word families in particular were extremely infrequent in at least one domain. A similar type of skewed distribution across disciplines was found in the case Gardner and Davies' AVL in a study conducted by Durrant (2016). He examined the coverage of this list in a corpus composed of assignments written by university students, and noted that the coverage of the list varied significantly across levels, genres (narrative recounting, literature review, etc.) and disciplines. Variation according to level can be desirable, since it might be useful in the evaluation of the quality of essays. However, the greatest amount of variation depended on the discipline considered. Such findings seem to go against the very viability of cross-disciplinary academic vocabulary lists. With this in mind, Hyland and Tse (2007) propose the creation of vocabulary lists specific to different scientific fields. Durrant (2016) is more optimistic and, noting the impracticability of highly field-specific language teaching, claims that there is a small core of the NAVL vocabulary (427 out of 3,000) that presents similar frequencies across genres.

The other main problem of academic lists, again according to Hyland and Tse (2007), is the lack of attention to the potential polysemy of vocabulary lists items. As a consequence, a given vocabulary item can display different central senses depending on the academic discipline in which it occurs (see Hyland and Tse, 2007, in their observations on the word families of *analysis*, *process* and *volume*). In this respect, it has been suggested that paying attention to lexical co-occurrence might be a way of dealing with the lack of information on meaning in traditional word lists (Hyland and Tse 2007; Hancioglu et al. 2008: 472-474).

The shortcomings pointed out by Hyland and Tse (2007) are particularly problematic for the use of academic lists in teaching: lemmas that are unproductive in a given field increase the learning burden without paying off; similarly, specific senses can be useful for students of a given domain, but not for others. When finding vocabulary for a writing aid, however, the burden of learning vocabulary somewhat unexploited in a given domain disappears: users are going to look up what they need. In any case, the cross-linguistic scope is still interesting in this scenario. As Frankenberg et al. (2018) point out, while it is expected for students of a given field to master technical vocabulary of that field, cross-disciplinary vocabulary (and combinations thereof) can go unnoticed, given the lack of specific focus on it.

The remainder of this paper explores the two discussed approaches in compiling word lists with a view to establishing which is best suited to our main goal, which is to identify the core vocabulary for an academic writing assistant. In this sense, the list of academic vocabulary is a starting point, and subsequently will be complemented with combinatorial information. Nevertheless, it must comprise a productive set of vocabulary that enters into characteristic combinations in academic discourse, rather than just a complement to general vocabulary enabling the user to understand academic texts. Although a keyword list seems from the start a more adequate option we will compare the outcomes of the subtractive and the keyword methods to confirm this hypothesis. Other methodological decisions, such as the way of establishing academic keyness and controlling for evenness of distribution, will be discussed as well. Finally, the productivity of the candidate lists in different domains will also be put to test by means of quantitative and qualitative analyses. More specifically, we will try to answer the following research questions:

RQ1. What are the consequences of opting for one of the two reviewed approaches (subtractive, keywords) in creating a list of academic Spanish vocabulary conceived as the core of an academic writing aid?

RQ2. To what extent such list can be considered cross-disciplinary, that is, to what extent its productivity is different depending on the academic field considered, both in terms of text coverage and the conveying of discipline-specific senses?

3. METHOD

The candidates to academic lemmas were extracted from a corpus containing 413 research articles. About half of the corpus came from the Spanish part of the *Spanish-English Research Articles Corpus* (SERAC; Pérez-Llantada, 2014). This corpus includes research articles published in indexed and peer-reviewed journals. This database was modified and complemented with other texts of the same type in order to obtain a more balanced sample in terms of the size of the four main domains of the corpus (see breakdown in Table 1). The size of the whole corpus is slightly over 2 million words. It is similar to the corpora used by Drouin (2007) and for the first version of French cross-disciplinary vocabulary (Tutin 2013), Coxhead (2000) or Paquot (2010). We made sure that the additional texts were research articles from indexed and/or peer-reviewed journals. Furthermore, the articles were originally published in pdf format and it was necessary to supervise their transfer to plain text.

The corpus is divided into four thematic sections: (i) Arts and Humanities (henceforth AH), (ii) Biological and Health Sciences (BHS), (iii) Physical Sciences and Engineering (PSE), and (iv) Social Sciences and Education (SOCS), each subdivided into three subsections, except for BHS, which is divided into two. This structure is drawn from SERAC. However, in contrast to SERAC, the four sections are balanced in terms of number of words rather than number of articles. Table 1 displays a breakdown of the corpus by sections:

DOMAIN	DISCIPLINE	no. of texts	words
AH (505,701 words)	Library Science	22	128,616
	Linguistics	30	204,245
	Literature	22	172,840
BHS (502,602 words)	Biology	46	206,011
	Medicine	98	296,591
PSE (506,644)	Physics/Chemistry	45	139,366
	Geology	28	154,967
	Engineering	54	212,311
SOCS (510,145)	Economy	22	138,366
	Education	25	154,868
	Sociology	22	216,911
Total	2,025,092		

Table 1. Breakdown of the academic corpus

As a contrast corpus, we used the fiction narrative part of the LEXESP corpus (Sebastián-Gallés et al., 2000), which contains ca. 5 million tokens. In this we follow Paquot (2010), who argues for the use of a strongly contrasting reference corpus.

Both corpora have been lemmatized and part-of-speech tagged with FreeLing (Padró and Stanilowsky, 2012).[1] This information has been used for the extraction of the vocabulary lists. Only lemmas corresponding to content words have been extracted, that is, nouns, adjectives, verbs and adverbs.

To ensure the specificity of the candidates of the keyword list (henceforth KWL), we compared the effects of two statistical tests: log-likelihood and t-test. The former is commonly used in corpora comparisons. The t-test is a parametrical one and therefore expects normally distributed data, which is not the case of corpus frequencies. Nevertheless, Paquot and Bestgen (2010) point out that this test is robust against non-normality violations. They also value positively the fact that

[1] The precision of FreeLing 3.0 PoS tagger for Spanish is 96.85% according to Gamallo et al. (2014). The narratives from LEXESP were re-annotated with the same version of FreeLing and the same parameters in order to make them comparable.

the t-test also seems to take into account the distribution of a word across corpora. For the t-test, we compared the frequencies normalized to 5,000 words in all the articles of the scientific corpus with frequencies in 5,000-word chunks of the narrative corpus. The values compared by means of log-likelihood were the total frequencies in each corpus. In all cases, the threshold for keyness was set at p ≤ .001. The t-test yielded a more conservative list than log-likelihood (1,223 as opposed to 7,379 candidate lemmas, out of a total of 35,698). Even if the t-test takes into account the distribution of candidates through corpus sections, preliminary inspections revealed the presence of specialized lexicon among the candidates selected through this method. This indicates the need for additional filters to control for evenness of distribution irrespective of the specificity measure chosen (see below).

In the case of the subtractive list (henceforth SpAWL), we simply excluded the lemmas occurring within the 2k most frequent ones in Davies (2006) from the candidates and applied other distributional criteria explained below.

Different methods were tested in order to control for the even distribution of the candidates. In the case of the keyword list produced using t-test, the combination with other filters produced too restrictive, scarcely productive lists (in terms of coverage), so that we proceeded only with the 7,379 lemmas yielded by log-likelihood.

Our first option was to filter our key academic lemmas with Gries's deviation of proportions (DP), which has several advantages as compared to the more traditional Juilliand D (see Gries 2008).[2] To calculate Gries' DP, we used the division into articles of the academic corpus. Restricting the initial candidates to the 20% with most homogeneous distributions in terms of DP and occurring in the four corpus sections yielded a list of 1,239 lemmas. This subset of key lemmas still included some with highly skewed distributions (up to a DP=0.96) and a preliminary inspection revealed the presence of intuitively specialized terms, such as *paciente* 'patient', *clínico* 'clinical', *capital* 'capital', *cardiovascular* 'cardiovascular', *hemoglobina* 'hemoglobine', etc. Raising DP's threshold had problematic consequences. For instance, using the DP of *paciente* 'patient' as cut-off point resulted in a 559 item list lacking interesting lemmas for academic production, such as *atención* 'attention', *evidencia* 'evidence', *probablity* 'probabilidad' etc. Finally, we decided to replace DP as a dispersion measure with two filters proposed by Gardner and Davies (2014) to control for evenness of distribution: lemmas should appear at least in 20% of texts and they should not have more than three times their expected frequency in each of the four sections of the corpus. This yielded a somewhat shorter list (833 after manually discarding some non-Spanish, erroneously lemmatized forms), but with a better coverage (see below).

In the case of SpAWL, the lemmas had to occur at least 10 times in all four main sections, and to be present in at least 7 of the 11 subsections. We did not use the absolute frequency filter

[2] The values of DP go from 0, which indicates no deviation with respect to the expected distribution per corpus section, to 1, which indicates that the frequency of the lemma in question deviates maximally from its expected distribution.

that Coxhead applied, given the shortness of the list resulting from the application of the two former criteria. In any case, given the first criterion, SpAWL candidates had a minimum frequency of 20 occurrences per million words.

Ideally, the repertoire included in academic lists should cover larger proportions in academic texts than in non-academic ones. Likewise, academic lemmas should not be specific of a particular scientific field, as noted above. To verify the extent to which the former assumptions hold, the coverage of the lists in the four subcorpora and in the contrast corpus was compared. Coverage was calculated using lemmas and part-of-speech (PoS) tags.

Likewise, to get an idea of the extent to which the list items are used across different disciplines with the same meanings, a study of the polysemy of a sample of the KWL list was conducted. This sample included the ten most common nouns of the list, given that frequent words seem to be more prone to polysemy (see Zipf 1936, and much more recently, Hernández-Fernández et al., 2016). For this sample, we examined lexical co-occurrent patterns by extracting two types of dependency relations: noun plus modifying adjective, and verb plus object noun. The corpus was parsed with UdPipe (Straka et al., 2016), a state-of-the-art neural network parser, which allowed for the extraction of these dependencies. The dependencies themselves had to be recurrent (adjectives and verbs had to occur at least 3 times with the target noun) and the lemmas in a given dependency relation shad to be significantly associated (a mutual information [MI] of ≥ 3). Thus, for instance, in the case of *trabajo* collocates such as *presente* 'this [paper]', *previo* 'previous', *reciente* 'recent', *numeroso* 'numerous', etc. suggested that it was used recurrently with the meaning 'piece of research', whereas collocates, such as *doméstico, femenino, masculino,* etc. indicated that it also conveyed the meaning 'work' (for a more detailed account of in which corpus sections these meanings were attested, see Section 4 below).

4. RESULTS AND DISCUSSION

Our first research question asked for the consequences of opting for either a subtractive method or keyword identification. The most evident consequence of this decision is the different size of the lists resulting from each method. The subtractive method, replicating that of Coxhead (2000), yields the shortest list, given that it excludes the 2k most frequent lemmas in Spanish (taken from Davies, 2006). Although its compilation is not strictly comparable to that of Coxhead's AWL, since her vocabulary units are word families and ours are lemmas, the fact that the Spanish list is much shorter is striking: in contrast to AWL's 570 word families, the Spanish list (SpAWL) contains only 389 lemmas (see Appendix 1). Precisely because word families are groupings of lemmas, one would expect a longer list in the Spanish case. However, the contrary is the case, probably because in Spanish there is a less accentuated gap between academic and non-academic lemmas (see below).

The KWL contains 833 lemmas (see Appendix 2). Since these lemmas, in contrast to those of SpAWL do not exclude frequent lemmas of general Spanish, it is interesting to examine the rate of overlapping between the Spanish KWL and the most frequent lemmas of Spanish. For this we have taken Davies (2006) as reference.

Davies (2006) bands	KWL	
	n	%
1+2K	637	76%
3K	83	10%
4K	32	4%
5K	17	2%
Out of the list	55-64	6-8%

Table 2. Coincidence between KWL and 5k most frequent lemmas of Spanish (Davies 2006)

The largest part of the KWL (76%) belongs to the 2k most frequent lemmas of general Spanish. Out of the remaining 196 (24%; see Appendix 3), 16% are among the bands of Davies' list going from 3K and 5K. Finally, between 6 and 8% of the lemmas are out of Davies' list (nine of these lemmas are multiwords and Davies' list does not contain this type of unit). This indicates that the KWL provides valuable indications as to the vocabulary that should be included in academic writing aid, since it is a summary of little over 800 lemmas of recurrent and cross-disciplinary academic vocabulary spread across the most frequent 5,000 lemmas of Spanish and beyond.

Due to its greater length and the fact that it contains highly frequent lemmas, the KWL should be much more productive than the SpAWL in terms of coverage. Table 3 compares the coverage of both lists and a fragment of the KWL with the same size of the SpAWL composed of the most frequent lemmas of the former.

	AH		BHS		PSE		SOCS		Narrative	
	tokens	%	tokens	%	tokens	%	tokens	%	tokens	%
SpAWL	19,863	4%	23,673	5%	26,179	5%	25,605	5%	70,852	1.5%
KWL	130,126	26%	128,846	26%	136,054	27%	148,153	29%	1,019,707	21%
Top 364 KWL	98,250	19%	101,895	20%	104,659	21%	112,204	22%	758,328	16%

total	505,701	502,602	506,644	510,145	4,833,249

Table 3. Coverage of the revised academic lists in academic and narrative corpora

These data confirm that the KWL is indeed much more productive than the SpAWL, not only due to its length, since the top 364 lemmas of the KWL have a much greater coverage than the SpAWL. On the other hand, the increase in coverage with respect to the narrative corpus is much more pronounced in the SpAWL, given the absence of frequent general vocabulary in this list.

In exchange, one would expect that, when the first 2K lemmas of Davies (2006) and the academic lists are combined, the increase of coverage in academic texts resulting from the addition of the 364 of the SpAWL would be much greater than that resulting from adding the 196 lemmas of the KWL. This is not completely confirmed, as shown in Table 4: the increase in coverage obtained from adding the academic lists is very similar (although somewhat larger in the case of the SpAWL), which means that the 196 entries of the KWL absent from Davies' list top entries have a productivity roughly comparable to that of the 364 SpAWL's items in academic discourse.

	AH		BHS		PSE		SOCS	
	tokens	%	tokens	%	tokens	%	tokens	%
Davies 2k	287,013	57%	265,341	53%	271,303	54%	297,567	58%
2k+SpAWL	306,876	61%	289,014	58%	297,482	59%	323,172	63%
2k+KWL	302,452	60%	286,335	57%	292,259	58%	318,108	62%

Table 4. Combined coverage of Davies' 2k most frequent word in Spanish and academic lists

The percentages corresponding to the 2k+SpAWL row are simply the result of adding the percentage covered by the two lists separately. Those corresponding to the 2k+KWL row are the percentage of texts covered by the union of Davies' 2k and the proposed keyword list. Although these coverage percentages cannot be compared in absolute terms to those given in Coxhead (2000)[3], the trend seems sufficiently clear: the KWL is more productive than the SpAWL.

These results are relevant to determine which method is better for compiling an academic list. This depends largely on the assumptions about academic vocabulary learning and use and on the purposes of the list itself. If academic vocabulary is conceived of as a sort of appendage to be added to a repository of frequent vocabulary, a list à la Coxhead could be the method of choice,

[3] Davies' 2K most frequent lemmas cover lower percentages of texts than the GSL first 2K word families, according to Coxhead's data. This probably has to do with the use of PoS tags (e.g. the coverage of the lemma *sobre,* which groups at least a noun 'envelope' and a preposition 'on', will be larger than two lemmas with the corresponding PoS tags). Thus, if we calculate the amount of academic text covered by these 2k lemmas without PoS tags, this yields a percentage of 73%, much more in accordance with Coxhead's (2000) results (75% of academic texts were covered by GSL).

especially for comprehension purposes: it is shorter than a keyword list, and the gains in coverage are somewhat larger.

However, the contribution of the SpAWL is considerably lower than that of its English counterpart (between 4% and 5% as opposed to a range between 9% to 12%). Whereas part of the differences can be due to the use of PoS, a phenomenon observed by Cobb and Horst (2004) could be at play here as well. They suggest that the gap between academic and non-academic vocabulary in languages other than English is less pronounced. They argue that the specialization of greco-latin terms for academic discourse is not so evident in other languages, either because they adapt their own heritage lexicon to academic needs (as in Dutch) or because greco-latin vocabulary is also pervasive in non-academic domains (as in French). In fact, Cobb and Horst (2004), analysing a subtractive academic list of French, voice their reservations regarding the need of such a resource, given that the first 2k lemmas of French cover similar percentages of academic texts as their English counterpart complemented with academic vocabulary.[4]

The evidence presented so far suggests that academic lists consisting of keywords are the only possible solution for languages like Spanish, even more so if the list's aims are production-oriented. Here a balance must be found between vocabulary specificity and productivity. That is, even if a lemma is highly frequent in general Spanish, it should be included in the academic list if it is also frequent in academic discourse. This has a double justification. First, as noted by Gilquin et al. (2007), it cannot be taken for granted that the more productive senses of academic lemmas are the same as in non-academic texts (see below the case of *trabajo* 'work'*,* which in the academic corpus is mostly used with the sense 'piece of research'). Second, the lemmas of the list could intervene in combinations that are exclusive of academic texts.

Only a cursory comparison between the SpAWL and the KWL gives us an idea of the potential loss in productivity resulting from using a subtractive approach. Among the nouns present in the KWL and absent from the SpAWL due to their being common in general Spanish one finds *decisión* 'decision', *estudio* 'study' or *trabajo* 'work, piece of research', piece of research', which occur in collocations crucial to express research-oriented or text-oriented contents: e.g. *decisión* is a recurrent object of *tomar* (MI=5.69) 'to make [a decision]' and a modifier of *toma* 'making [of a decision]' (MI=6.30)—both seemingly research-oriented collocations. Several of the collocations of *estudio* and *trabajo* can be used to provide textual or evidential (source) indications (*estudio presente* (MI=3.75)*/precedente* (MI=3.71)*/reciente* (MI=3.59) 'this/previous/recent study'; *trabajo anterior/previo* 'previous piece of research'), while others seem related to conveying research-oriented meanings, like the topic or the aim of the research (e.g. *trabajo* as subject of *analizar* (MI=1.3) 'to analyze' and *pretender* (MI=3.8) 'to intend'). Some of these combinations are

[4] In connection with this, see Goodfellow et al. (2002), whose results show that the 2k most frequent French lemmas are better predictors of essay quality than a subtractive list of academic terms.

characteristic to academic discourse. Thus, *trabajo* as subject of *analizar* does not seem possible when the former expresses the sense ‘job’, rather than ‘piece of research’.

Turning now to RQ2, it asked to what extent cross-linguistically academic vocabulary had the same productivity in different academic fields. Coverage data in Tables 3 and 4 provide part of the answer. The two academic lists alone cover larger percentages of text in the SOCS and PSE than in the AH and BHS subcorpora. On the other hand, when the academic and the general 2k most frequent lemmas are combined, the AH and SOCS are the best-covered subcorpora. This could mean that AH texts use specifically academic vocabulary less frequently than the other three fields. In contrast, SOCS is in an intermediate position in the sense that academic vocabulary is here more productive than in other fields, but also is general vocabulary. BHS and PSE subcorpora use academic vocabulary more frequently than AH and general vocabulary less frequently than both AH and SOCS, which suggests that a larger proportion of these texts is covered by specific, non-cross-disciplinary technical vocabulary.

A more complete picture can be obtained by exploring the different productivity of KWL lemmas across the four subcorpora. Table 5 displays the top ranked lemmas according to their frequency, those in middle positions and those ranked the lowest for the four subcorpora.

	AH		BHS		PSE		SOCS	
rank	lemma	PoS	lemma	PoS	lemma	PoS	lemma	PoS
1	ser	V	ser	V	ser	V	ser	V
2	no	R	haber	V	haber	V	no	R
3	haber	V	no	R	poder	V	haber	V
4	más	R	estudio	N	no	R	poder	V
5	poder	V	más	R	más	R	más	R
6	tener	V	mayor	A	valor	N	tener	V
7	estar	V	poder	V	obtener	V	trabajo	N
8	decir	V	realizar	V	estar	V	estar	V
9	también	R	caso	N	utilizar	V	social	A
10	hacer	V	grupo	N	realizar	V	forma	N
412	relativoA		fecha	N	cercano	A	etapa	N
413	segundo	N	final	A	continuación	N	final	A
414	fuerte	A	literatura	N	llevar	V	comprobar	V
415	utilización	N	utilización	N	relativamente	R	necesitar	V
416	difícil	A	asimismo	R	ambiental	A	común	A
417	herramienta	N	final	N	básico	A	cualitativo	A
418	nacional	A	por_ejemplo	R	global	A	depender	V
419	visión	N	derivar	V	implicar	V	estado	N
420	claramente	R	efectuar	V	recoger	V	incorporar	V
421	lograr	V	limitación	N	sitio	N	método	N
824	tasa	N	volver	V	personal	A	cálculo	N
825	disminución	N	comentar	V	historia	N	colocar	V
826	intervalo	N	entrar	V	participar	V	disminución	N
827	lineal	A	teoría	N	opinión	N	experimental	A

828	longitud	N	desear	V	excluir	V	localizar	V
829	ambiental	A	exigir	V	pregunta	N	sitio	N
839	temperatura	N	opinión	N	sociedad	N	posiblemente	R
831	coeficiente	N	partir	V	informar	V	curva	N
832	curva	N	precisamente	R	sexo	N	temperatura	N
833	muestreo	N	interesar	V	mujer	N	longitud	N

Table 5. Highest-, middle- and lowest-ranked lemmas by their frequency in the four subcorpora

The top ranked lemmas for each subcorpus are similar. They tend to be lemmas with very general meanings and potentially polysemous (*ser* ‘to be’, *tener* ‘to have’, *más* ‘more’). However, among these top-ranked lemmas some differences can be already perceived: all but the AH subcorpus include some content words, seemingly relative to research procedures (*realizar* ‘to carry out’ in BHS and PSE, *caso* ‘case’ in BHS) or research topics (*valor* ‘value’ in PSE*, social* ‘social’ in SOCS). The least frequent lemmas show similarities between AH and SOCS, on the one hand (mathematics related terms, such as *curva* ‘curve’ and *longitud* ‘length’ in both subcorpora, *coeficiente* ‘coefficient’ in AH, *cálculo* ‘calculus’ in SOCS), and BHS and PSE, on the other (seemingly evaluative expressions, such as *opinión* ‘opinion’ in both subcorpora, *interesar* ‘to interest’ or *desear* ‘to desire’ in BHS). Furthermore, the social-related lemmas occurring at the end of the PSE ranking are much more frequent in the SOCS subcorpus: in the latter, *mujer* ‘woman’ appears in the position 37 and *society* ‘society’ in the position 67. From this evidence, it can be hypothesized that the most frequent KWL lemmas are similarly used in all four fields, while somewhere in the middle of the list they start to display different degrees of usefulness depending on the discipline considered.

A quantitative exploration can offer a still more precise view of the differences in productivity of academic lemmas across fields. For this, we ranked the lemmas of the whole KWL by their frequency in the four subcorpora and compare the correlations of these rankings. Table 6 shows Spearman correlation coefficients between these four rankings, plus the ranking resulting from the whole corpus’ frequencies.

	Whole corpus	AH	BHS	PSE	SOCSC
Whole corpus	1	.75	.73	.74	.80
AH		1	.33	.38	.76
BHS			1	.65	.40
PSE				1	.40
SOCSC					1

Table 6. Rank correlations for the KWL between the whole academic corpus and its four fields

The most similar orderings are those corresponding to the AH and SOCSC subcorpora (.76) followed by BHS and PSE (.65). The correlations between AH, on the one hand, and BHS and PSE, on the other, are low, as are those between the latter two disciplines and SOCS. This again confirms a divide between soft and hard sciences, like the coverage data seen in Table 4. Nevertheless, it should be remembered that the lemmas of the KWL have a relatively homogeneous distribution across the four subcorpora. Likewise, the correlations between the ranking of the whole corpus and those of the four subcorpora are similarly high. These two facts argue for a similar potential interest of the whole list for all four scientific domains, even if some of its lemmas are exploited differently depending on the discipline.

In addition to the differing productivity of academic lemmas across scientific fields, Hyland and Tse (2007) see it as problematic that a given lemma can convey different meanings in different disciplines. The examination of co-occurrence data of the ten most frequent nouns from the KWL contradicts to some extent Hyland and Tse's results. Out of the ten nouns examined, only five expressed clearly field-specific senses: *estudio* was repeatedly used with the meaning of 'formal education' in the SOCS subcorpus; *caso* meant 'disease or injury' in the BHS subcorpus; *valor* was used with the sense of 'moral principles' in SOCS; *forma* occurred repeatedly as 'shape' in BHS; and finally *trabajo* regularly meant 'work' in the SOCS subcorpus. However, along with these subcorpus-specific senses, these lemmas also displayed others common to the whole corpus: *estudio* 'study, piece of research' and 'process of research'; *caso* 'set of circumstances'; *valor* 'value, quantity'; trabajo 'study, piece of research', and *forma* 'way, manner'. The case of the noun *análisis*—equivalent to one of those studied by Hyland and Tse—is difficult to interpret, since it can be debated whether it has a general meaning ('detailed examination') employed in the whole corpus, or if it is used as a part of multiword technical terms (*análisis granulométrico / cuantitativo / estadístico...* 'granulometric/quantitative/statistical analysis). Finally, four nouns (*resultado* 'result', *tipo* 'type, kind', *nivel* 'level', *parte* 'part') seemed to display only shared senses across different domains.

5. CONCLUSIONS

This article has studied the process of compiling an academic vocabulary list for Spanish. Several methods have been tested and two lists have been proposed: one obtained through what we called subtractive approach and the other composed of key lemmas of an academic corpus. In both cases, we have employed filters to ensure that candidate lemmas were evenly distributed across the subsections of the academic corpus.

After comparing the two lists, we have argued that the subtractive approach, which can be an economic resource for comprehension purposes in a language like English, is problematic for Spanish data. The list resulting from following this approach was shorter and less productive than its English counterpart. It has been hypothesised that this is the result of Spanish having a less pronounced gap than English between general and academic vocabularies, at least as far as lemmas (and not senses nor word combinations) are concerned.

Keyword lists are more adequate for productive tasks and resources focused on such tasks, due to their greater autonomy, in comparison to subtractive lists: they do not depend on the existence of other repertoires accounting for frequent general vocabulary. In this regard, the KWL contains frequent lemmas of general Spanish that are also very productive in academic texts. Furthermore, those lemmas can convey senses specific of academic discourse (e.g. *trabajo* 'piece of research'). Finally, given that the final purpose of the proposed list is to provide academic vocabulary for a writing aid, suppressing frequent lemmas would have meant to renounce to productive combinations in academic discourse.

When it comes to the productivity of the list in different academic domains, a hiatus has been observed between soft and hard sciences. Taken individually, there are lemmas that probably have different usability degrees depending on the discipline. However, the list as a whole seems to be a good compromise solution, as the even distribution of its lemmas has been controlled for. Likewise, judging after the most frequent noun lemmas, it seems that list items share meanings across the four subcorpora, even if along with those shared meaning, they convey some domain-specific senses.

Finally, it should also be noted that this list has not been conceived of as a stand-alone product, but as the starting point for more sophisticated resources that will include semantic and combinatorial information. The extraction of such information will be the object of future research. Given that the proposed list has been compiled as a resource to identify the core vocabulary of a writing assistant, it could be also interesting to study its applicability to domains such as the teaching of Spanish for specific purposes or text quality assessment, topics that fall outside the scope of the present paper. In any case, the present proposal fills a gap in the study of the vocabulary of academic Spanish, which, in sharp contrast to other languages, has hitherto lacked a repertoire such as the one provided here.

REFERENCES

Ackermann, K., & Chen, Y. H. (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

Almela, R., Cantos, P., Sánchez, A., Sarmiento, R. & Almela, M. (2005). *Frecuencias del español. Diccionario y estudios léxicos y morfológicos*. Madrid: Editorial Universitas.

Baptista, J., Costa,N., Guerra, I., Zampieri, M., Cabral, M. & Mamede, N. (2010). "P-AWL: Academic Word List for Portuguese", in T.A.S. Pardo, A. Branco, A. Klautau, R. Vieira, V.L.S. de Lima (eds), *Computational Processing of the Portuguese Language. PROPOR 2010. Lecture Notes in Computer Science*, vol 6001. Berlin, Heidelberg: Springer.

Bogaards, P. (2001). “Lexical Units and the Learning of Foreign Language Vocabulary”, *Studies in Second Language Acquisition* 23/3, 321–343. https://doi.org/10.1017/S0272263101003011

Buchanan, M. (1927). *A Graded Spanish Word Book*. Toronto: Toronto University Press.

Capel, A. (2012). “Completing the English Vocabulary Profile: C1 and C2 vocabulary”, *English Profile Journal* 3: e1. https://doi.org/doi:10.1017/S2041536212000013

Cobb, T., & Horst, M. (2004). “Is there room for an academic word list in French?”, in P. Bogaards and B. Laufer (eds.), *Vocabulary in a Second Language. Selection, acquisition, and testing*. Amsterdam/Philadelphia: John Benjamins: 15–38.

Coxhead, A. (2000). “A new academic word list”, *TESOL Quarterly* 34/2, 213–238.

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.2018.02.00

Davies, M. (2006). *Frequency Dictionary of Spanish*. New York: Routledge.

Drouin, P. (2007). “Identification automatique du lexique scientifique transdisciplinaire”, *Revue Française de Linguistique Appliquée* 12/2, 45–64.

Durrant, P. (2016). ‘To what extent is the Academic Vocabulary List relevant to university student writing?’, *English for Specific Purposes* 43, 49–61. http://doi.org/10.1016/j.esp.2016.01.004

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P. & Sharma, N. (2018). “Developing a writing assistant to help EAP writers with collocations in real time”, *ReCALL First View,* 1–17, doi: 10.1017/S0958344018000150

Gamallo, P., Pichel, J. C., & Garcia, M. (2014). “Análisis morfosintáctico y clasificación de entidades nombradas en un entorno Big Data”, *Procesamiento Del Lenguaje Natural 53*, 17–24.

García Hoz, V. (1953). *Vocabulario común, vocabulario usual y vocabulario fundamental*. Madrid: CSIC.

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”.Iin N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani,

H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 260-265, Miyazaki, Japan, 2018.

Gardner, D., & Davies, M. (2014). “A new academic vocabulary list”, *Applied Linguistics* 35/3, 305–327. https://doi.org/10.1093/applin/amt015

Gilquin, G., Granger, S. & Paquot, M. (2007). “Learner corpora: The missing link in EAP pedagogy”, *Journal of English for Academic Purposes* 6/4: 319–335. https://doi.org/10.1016/j.jeap.2007.09.007

Goodfellow, R., Lamy, M.-N., & Jones, G. (2002). “Assessing learners’ writing using lexical frequency”, *ReCALL* 14/01, 133–145. https://doi.org/10.1017/S0958344002001118

Granger, S., & Paquot, M. (2015). “Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid”, *Lexicographica: International Annual for Lexicography* 31/1, 118–141.

Gries, S. T. (2008). “Dispersions and adjusted frequencies in corpora”, *International Journal of Corpus Linguistics* 13/4, 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Hancioğlu, N., Neufeld, S., & Eldridge, J. (2008). “Through the looking glass and into the land of lexico-grammar”, *English for Specific Purposes* 27/4, 459–479. https://doi.org/10.1016/j.esp.2008.08.001

Hyland, K., & Tse, P. (2007). “Is there an “academic vocabulary”?”, *TESOL Quarterly* 41/2, 235–253. https://doi.org/10.1002/j.1545-7249.2007.tb00058.x

Jacques, M. P., & Tutin, A. (2018). “Introduction”, in A. Tutin, and M.-P. Jacques (eds.), *D’une discipline à l’autre. Lexique transversal et formules discursives des sciences humaines*. London: ISTE.

Johansson Kokkinakis, S., Sköldberg, E., Henriksen, B., Kinn, K., & Johannessen, J. B. (2012). “Developing Academic Word Lists for Swedish, Norwegian and Danish – a joint research project”, in R.V. Fjeld, and J. M. Torjusen (ed.), *Proceedings of the 15th EURALEX International Congress.* Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, 563–569.

Juilland, A. G., & Chang-Rodríguez, E. (1964). *Frequency Dictionary of Spanish Words*. The Hague: Mouton.

Keniston, H. (1920). “Common words in Spanish”, *Hispania* 3, 242–250.

Lafon, P. (1980). “Sur la variabilité de la fréquence des formes dans un corpus”, *MOTS* 1, 128-165.

Laufer, B., & Nation, P. (1995). “Vocabulary Size and Use: Lexical Richness in L2 Written Production”, *Applied Linguistics* 16/3, 307–322. https://doi.org/10.1093/applin/16.3.307

Lea, D. (2014). “Making a Learner’s Dictionary of Academic English”, in A. Abel, C. Vettori, & N. Ralli (eds.), *Proceedings of the 16th EURALEX International Congress*. Bolzano: EURAC research, 181–189.

Montolío, E. (2014), *Manual de escritura académica y profesional*. Barcelona: Ariel.

Nation, I. S. P., & Xue, G. (1984). “A university word list”, *Language Learning and Communication* 3, 215–229.

Padró, L. & Stanilovsky, E. (2012). “FreeLing 3.0: Towards Wider Multilinguality”, in N. Carzolari et al. (eds.), *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA): 2473-2479.

Paquot, M. (2010). *Academic vocabulary in learner writing*. London: Continnuum. https://doi.org/10.1007/s13398-014-0173-7.2

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.

Phal, A. (1971). *Vocabulaire général d’orientation scientifique (VGOS). Part du lexique commun dans l’expression scientifique*. Paris: Didier.

Rodríguez Bou, I. (1952). *Recuento de vocabulario español*. Puerto Rico: Universidad de Puerto Rico.

Sebastián-Gallés, N., Martí Antonín, M.A., Carreiras Valiña, M. F., & Cuetos Vega, F. (2000). *LEXESP: Léxico informatizado del español*. Barcelona: Edicions de la Universitat de Barcelona.

Straka, M., J. Hajic, & Straková, J. (2016). ‘Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing’, in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA): 1659-1666.

Tutin, A. (2013). “Lexique transdisciplinaire (V1)”. https://scientext.hypotheses.org/lexique-transdisciplinaire-v1 [retrieved 3.2019].

Verdaguer, I., N. J. Laso, T. Guzmán-González, D. Salazar, E. Comelles, E, Castaño,and J. Hilferty. 2013. “Scie-Lex. A lexical database”, in I. Verdaguer, N. J. Laso, and D. Salazar (eds.), *Biomedical English: A corpus-based approach*. John Benjamins: Amsterdam/Philadelphia, 21–38.

Appendix 1. Spanish Academic Word List (SpAWL)

Nouns

actualidad, adaptación, administración, ahorro, algoritmo, apartado, apertura, aplicación, aporte, aproximación, asignación, asociación, barrera, bloque, caracterización, ciclo, clasificación, cobertura, código, colaboración, combinación, complejidad, complejo, componente, conexión, consumo, contexto, contrario, contraste, contribución, corrección, correlación, correspondencia, coste, cumplimiento, década, dependencia, descripción, desviación, determinación, difusión, diseño, disponibilidad, dispositivo, distribución, diversidad, duración, eficacia, eficiencia, eje, elaboración, eliminación, enfoque, entidad, entorno, esquema, establecimiento, estadística, estándar, estrategia, evaluación, evento, evidencia, expansión, explotación, extensión, fase, finalidad, flujo, fragmento, funcionamiento, fundamento, gestión, guía, herramienta, hipótesis, identificación, impacto, implantación, implicación, incidencia, inclusión, incorporación, incremento, indicador, inicio, inserción, instancia, intensidad, interacción, introducción, inversión, investigador, limitación, localización, magnitud, manejo, mantenimiento, mapa, matriz, media, mejora, metodología, modificación, núcleo, opción, orientación, parámetro, pauta, perfil, periodo, plataforma, porcentaje, potencial, predominio, probabilidad, procesamiento, productividad, promedio, proporción, propuesta, publicación, puesta, rango, realización, recogida, recorrido, reducción, referente, registro, regulación, relevancia, rendimiento, repercusión, requisito, resolución, responsable, restricción, resumen, revisión, rol, secuencia, segmento, seguimiento, selección, separación, similitud, síntesis, software, soporte, sugerencia, tecnología, test, toma, totalidad, transformación, transición, transmisión, ubicación, umbral, unión, utilidad, utilización, validación, validez, valoración, variabilidad, variable, variación, varianza, variedad

Adjectives

aceptable, adicional, adulto, alternativo, analítico, básico, biológico, característico, cierto, comparativo, complementario, considerable, consistente, convencional, creciente, cualitativo, cuantitativo, dependiente, descriptivo, determinante, digital, disponible, dominante, electrónico, específico, estadístico, estándar, estricto, estructural, exclusivo, existente, experimental, extenso, externo, favorable, funcional, geográfico, global, gran, heterogéneo, idéntico, imprescindible, indirecto, industrial, intermedio, mediano, mismo, occidental, parcial, pasivo, perteneciente, potencial, predominante, preliminar, procedente, productivo, progresivo, proveniente, regional, relevante, representativo, restante, resultante, secundario, selectivo, sistemático, tecnológico, temporal, teórico, urbano, válido, variable, visible, visual

Verbs

abarcar, abordar, activar, adaptar, adecuar, agrupar, aislar, apropiar, aproximar, asignar, brindar, caracterizar, centrar, clasificar, combinar, complementar, comportar, comprometer, condicionar, conllevar, constatar, contrastar, corroborar, cuantificar, desempeñar, detallar, detectar, diferenciar, diferir, dificultar, diseñar, efectuar, enfocar, englobar, especializar, especificar, estructurar, evaluar, evidenciar, extraer, favorecer, finalizar, financiar, formar_parte, garantizar, generalizar, generar, hacer_referencia, incrementar, inducir, influir, involucrar, ligar, llevar_a_cabo, mediar, moderar, motivar, optar, orientar, originar, oscilar, poner_de_manifiesto, predominar, prever, privar, promover, reforzar, regular, resaltar, restringir, resumir, seleccionar, situar, tener_en_cuenta, valorar, variar, verificar, vincular

Adverbs

actualmente, adecuadamente, altamente, ampliamente, anteriormente, aproximadamente, asimismo, bastante, cuanto, en_consecuencia, en_gran_medida, específicamente, estadísticamente, frecuentemente, fuertemente, fundamentalmente, habitualmente, independientemente, inicialmente, ligeramente, más, mayoritariamente, mejor, mucho, parcialmente, particularmente, poco, por_consiguiente, por_ejemplo, por_el_contrario, por_lo_tanto, por_otro_lado, posteriormente, previamente, primero, principalmente, recientemente, relativamente, respectivamente, solo, tanto

Appendix 2. Academic Spanish Keyword List

Nouns

acceso, acción, actividad, actualidad, acuerdo, adaptación, agente, agua, alternativa, ambiente, ámbito, análisis, año, aparición, apartado, aplicación, apoyo, área, artículo, asociación, aspecto, atención, aumento, ausencia, autor, ayuda, base, beneficio, búsqueda, cálculo, calidad, cambio, campo, cantidad, capacidad, carácter, característica, carga, caso, categoría, causa, centro, ciclo, circunstancia, clase, clasificación, clave, coeficiente, colaboración, combinación, comparación, componente, comportamiento, composición, comunicación, comunidad, concentración, concepto, conclusión, condición, confianza, conjunto, conocimiento, consecuencia, consideración, construcción, consumo, contacto, contenido, contexto, continuación, contribución, control, correlación, corte, creación, crecimiento, criterio, cuadro, cuerpo, cuestión, curva, dato, década, decisión, definición, demanda, desarrollo, descripción, desviación, determinación, día, diferencia, dificultad, dimensión, dirección, diseño, disminución, distancia, distribución, diversidad, duda, edad, efecto, eficacia, ejemplo, elaboración, elección, elemento, embargo, empleo, enfoque, entorno, equipo, error, escala, esfuerzo, espacio, esquema, estado, estrategia, estructura, estudio, etapa, evaluación, evidencia, evolución, existencia, éxito, experiencia, explicación, expresión, extensión, factor, falta, familia, fase, fecha, fenómeno, figura, fin, final, finalidad, flujo, fondo, forma, formación, frecuencia, fuente, fuerza, función, funcionamiento, futuro, generación, general, género, *grado, grupo, hecho, herramienta, hipótesis, historia, hombre, hora, idea, identificación, imagen, impacto, implicación, importancia, inclusión, incremento, indicador, índice, individuo, influencia, información, inicio, institución, instrumento, interacción, interés, interpretación, intervalo, intervención, investigación, investigador, lado, lectura, limitación, límite, línea, literatura, longitud, lugar, luz, manejo, manera, marco, masa, material, mayoría, mecanismo, media, medida, medio, mejora, mercado, método, metodología, miembro, mitad, modelo, modificación, modo, momento, motivo, movimiento, muestra, muestreo, mujer, mundo, naturaleza, necesidad, nivel, nombre, norma, número, objetivo, objeto, observación, obtención, ocasión, opción, operación, opinión, oportunidad, orden, orientación, origen, país, papel, par, parámetro, parte, paso, patrón, pérdida, perfil, periodo, período, persona, perspectiva, peso, población, poder, porcentaje, posibilidad, posición, potencial, práctica, pregunta, presencia, presión, principio, probabilidad, problema, procedimiento, proceso, producción, producto, programa, promedio, propiedad, proporción, propósito, propuesta, proyecto, prueba, puesto, punto, rango, rasgo, razón, reacción, realidad, realización, recurso, red, reducción, referencia, región, registro, relación, relevancia, representación, resolución, responsable, respuesta, resto, resultado, resumen, revisión, riesgo, salud, sección, sector, secuencia, seguimiento, segundo, selección, sentido, serie, servicio, sexo, sistema, sitio, situación, sociedad, solución, sujeto, tabla, tamaño, tarea, tasa, técnica, tema, temperatura, tendencia, teoría, término, test, tiempo, tipo, toma, total, totalidad, trabajo, transformación, tratamiento, unidad, uso, utilidad, utilización, valor, valoración, variable, variación, variedad, ventaja, versión, vez, vía, vida, visión, vista, volumen, zona*

Adjective

1/uno, 2/dos, 3/tres, activo, actual, alto, ambiental, amplio, anterior, anual, bajo, básico, bueno, capaz, central, cercano, científico, cierto, claro, complejo, completo, común, concreto, constante, continuo, correcto, correspondiente, corto, crítico, cualitativo, cuantitativo, descriptivo, diferente, difícil, directo, disponible, distinto, diverso, doble, económico, efectivo, escaso, esencial, español, especial, específico, estadístico, estándar, estructural, evidente, existente, experimental, externo, fácil, final, físico, frecuente, fuerte, funcional, fundamental, futuro, general, geográfico, global, gran, grande, habitual, humano, igual, importante, independiente, individual, inferior, inicial, interesante, internacional, interno, largo, libre, lineal, local, máximo, mayor, medio, mejor, menor, mínimo, mismo, múltiple, nacional, natural, necesario, negativo, normal, nuevo, numeroso, original, parcial, particular, pequeño, personal, perteneciente, posible, positivo, posterior, potencial, práctico, preciso, presente, previo, primario, principal, profundo, propio, próximo, público, rápido, real, reciente, relativo, relevante, representativo, restante, secundario, significativo, siguiente, similar, simple, social, solo, suficiente, superior, técnico, temporal, teórico, típico, total, tradicional, último, único, útil, variable

Verbs

abarcar, abordar, abrir, aceptar, acompañar, actuar, adaptar, adecuar, adoptar, adquirir, afectar, afirmar, agradecer, agrupar, aislar, ajustar, alcanzar, analizar, añadir, aparecer, aplicar, aportar, apoyar, apreciar, apropiar, aproximar, apuntar, asegurar, asignar, asociar, asumir, atender, atribuir, aumentar, avanzar, ayudar, basar, buscar, caber, calcular, cambiar, caracterizar, causar, centrar, cerrar, citar, clasificar, coincidir, colocar, combinar, comentar, comenzar, comparar, compartir, completar, componer, comprender, comprobar, concluir, condicionar, conducir, confirmar, conformar, conocer, conseguir, considerar, consistir, constituir, construir, contar, contemplar, contener, continuar, contribuir, controlar, convertir, corresponder, crear, creer, cubrir, cumplir, dar, deber, decidir, decir, dedicar, definir, dejar, demostrar, denominar, depender, derivar, desarrollar, describir, desear, destacar, detallar, detectar, determinar, diferenciar, diferir, dirigir, diseñar, disminuir, disponer, distinguir, distribuir, dividir, efectuar, ejercer, elaborar, elegir, elevar, eliminar, empezar, emplear, encontrar, entender, entrar, esperar, establecer, estar, estimar, estudiar, evaluar, evidenciar, evitar, excluir, exigir, existir, experimentar, explicar, exponer, expresar, extender, extraer, facilitar, favorecer, figurar, fijar, formar, formar_parte, funcionar, garantizar, generar, haber, hablar, hacer, hacer_referencia, hallar, identificar, impedir, implicar, incluir, incorporar, incrementar, indicar, influir, informar, iniciar, integrar, intentar, interesar, interpretar, intervenir, introducir, ir, justificar, limitar, llamar, llegar, llevar, llevar_a_cabo, localizar, lograr, manifestar, mantener, marcar, medir, mejorar, mencionar, modificar, mostrar, necesitar, observar, obtener, ocupar, ocurrir, ofrecer, orientar, oscilar, parecer, participar, partir, pasar, pensar, perder, permanecer, permitir, pertenecer, plantear, poder, poner, poner_de_manifiesto, poseer, precisar, presentar, pretender, proceder, producir, promover, proponer, proporcionar, provocar, publicar, quedar, querer, realizar, recibir, recoger, recomendar, reconocer, recordar, reducir, referir, reflejar, relacionar, repetir, reportar, representar, requerir, resaltar, resolver, responder, resultar, resumir, revelar, revisar, saber, seguir, seleccionar, señalar, separar, ser, servir, significar, situar, soler, someter, suceder, sufrir, sugerir, superar, suponer, surgir, tender, tener, tener_en_cuenta, tomar, trabajar, transformar, tratar, ubicar, unir, usar, utilizar, valorar, variar, venir, ver, vincular, vivir, volver

Adverbs

actualmente, además, ahora, anteriormente, antes, aproximadamente, aquí, así, asimismo, aun, bastante, bien, casi, cerca, claramente, directamente, entonces, especialmente, estadísticamente, exclusivamente, finalmente, fundamentalmente, generalmente, hoy, igualmente, incluso, luego, más, mejor, menos, mucho, muy, no, poco, por_ejemplo, por_el_contrario, por_lo_tanto, por_otro_lado, posiblemente, posteriormente, precisamente, previamente, primero, principalmente, probablemente, relativamente, respectivamente, sí, siempre, significativamente, solamente, solo, sólo, también, tampoco, tan, tanto, todavía, únicamente, ya

Appendix 3. Keyword list’s lemmas outside the most frequent 2k lemmas of Spanish (n=196)

Nouns

actualidad, adaptación, apartado, aplicación, asociación, ciclo, clasificación, coeficiente, colaboración, combinación, componente, consumo, contexto, contribución, correlación, curva, década, descripción, desviación, determinación, diseño, disminución, distribución, diversidad, eficacia, elaboración, enfoque, entorno, esquema, estrategia, evaluación, evidencia, extensión, fase, finalidad, flujo, funcionamiento, herramienta, hipótesis, identificación, impacto, implicación, inclusión, incremento, indicador, inicio, interacción, intervalo, investigador, limitación, longitud, manejo, media, mejora, metodología, modificación, muestreo, obtención, opción, orientación, parámetro, perfil, periodo, porcentaje, potencial, probabilidad, promedio, proporción, propuesta, rango, realización, reducción, registro, relevancia, resolución, responsable, resumen, revisión, secuencia, seguimiento, selección, tasa, test, toma, totalidad, transformación, utilidad, utilización, valoración, variable, variación, variedad

Adjectives

1/uno, 2/dos, 3/tres, ambiental, anual, básico, cierto, cualitativo, cuantitativo, descriptivo, disponible, específico, estadístico, estándar, estructural, existente, experimental, externo, funcional, geográfico, global, gran, lineal, mismo, parcial, perteneciente, potencial, relevante, representativo, restante, secundario, temporal, teórico, variable

Verbs

abarcar, abordar, adaptar, adecuar, agrupar, aislar, apropiar, aproximar, asignar, caracterizar, centrar, clasificar, combinar, condicionar, detallar, detectar, diferenciar, diferir, diseñar, efectuar, evaluar, evidenciar, excluir, extraer, favorecer, formar_parte, garantizar, generar, hacer_referencia, incrementar, influir, llevar_a_cabo, localizar, orientar, oscilar, poner_de_manifiesto, promover, reportar, resaltar, resumir, seleccionar, situar, tener_en_cuenta, valorar, variar, vincular

Adverbs

actualmente, anteriormente, aproximadamente, asimismo, bastante, estadísticamente, fundamentalmente, más, mejor, mucho, poco, por_ejemplo, por_el_contrario, por_lo_tanto, por_otro_lado, posteriormente, previamente, primero, principalmente, relativamente, respectivamente, significativamente, solo, tanto