Exploiting a Corpus to Compile a Lexical Resource for Academic Writing: Spanish Lexical Combinations

Margarita Alonso-Ramos, Marcos García-Salido and Marcos Garcia

Universidade da Coruña, Grupo LyS
Departamento de Letras, Facultade de Filoloxía
E-mail: lxalonso@udc.es, marcos.garcias@udc.gal, marcos.garcia.gonzalez@udc.gal

Abstract

This paper provides insight into ongoing research focusing on the exploitation of Spanish academic corpora in order to build up a lexical tool addressed to novice writers of academic texts. The object of the lexical tool is what we call *academic lexical combinations* (ALC). By ALC we mean recurrent segments of words that may or may not be semantically compositional and fulfill rhetorical functions such as giving examples, concluding, expressing emphasis, etc. These functions are particularly prominent in academic discourse. ALCs comprise from collocations to idioms as well as formulas, as they are understood in the Meaning-Text Theory (Mel'čuk, 2012). The procedure adopted for the extraction of the ALC from the corpus is described along with how we combine statistical information and native speakers' intuition. Even if corpora play a leading role in the construction of our lexical tool, we need to filter out corpus output with phraseological criteria, which makes human intervention necessary. Finally, we specify the architecture of the lexical tool and we show different prototype lexicographical entries.

Keywords: academic language; collocation; idiom; formula; lexical bundle; corpus

1. Introduction

In today's knowledge society, the written text plays a primary role, especially in the academic context. When students get into the university, they have to face a new discourse genre and need tools that allow them not only to understand academic texts, but also to produce them. Whereas languages such as English are relatively well provided in this respect (McCarthy & O'Dell, 2008; Swales & Feak, 2012; Lea et al., 2014), no resource of this kind exists for Spanish so far. Although academic writing is a multifaceted phenomenon, we believe that the gist of acquiring academic writing skills resides in learning what we call *academic lexical combinations* (ALCs). By ALC we mean recurrent segments of words that may or may not be semantically compositional and that fulfill rhetorical functions such as giving examples, concluding, expressing possibility or certainty, etc. These functions are particularly prominent in academic discourse. ALCs comprise from collocations (*extraer conclusiones* 'to draw conclusions') to idioms (*en conclusió*n 'in conclusion') as well as formulas that traditionally do not have a place in the phraseological spectrum (roughly *lexical bundles*, Biber et al. (2004), as *como se ha dicho previamente* 'as stated previously').

The literature on English ALC is extensive, especially on lexical bundles (Biber et al., 1999, 2004; Cortes, 2004; Hyland, 2008; Verdaguer & Salazar, 2013; Salazar, 2014). The reason for this growing interest lies mainly in the predominant role of English as an international academic language. Therefore, there is a need to build up lexical resources helping principally non-native English speakers to write research articles or, more generally, academic texts. In recent years, some initiatives to compile academic lexical resources in languages other than English have been undertaken as well. Though the following is not an exhaustive list, we can mention some projects on European languages. For French an extensive academic corpus has been compiled around the project Scientext, which has served as the basis for a considerable amount of research into French phraseology (Tutin & Grossmann, 2014; Cavalla & Loiseau, 2014; Tutin, 2010, 2014). Likewise, academic Brazilian Portuguese, especially that found in article abstracts, has been the focus of a research

team based at the Universidade do Rio Grande do Sul (Krause Kilian & Dias Loguercio, 2015). Similar projects in other languages are less advanced. For instance, there is a joint multi-disciplinary Scandinavian project aimed at developing three new academic lexical resources based on corpora consisting of texts from Swedish, Norwegian and Danish academic settings (Johansson Kokkinakis et al., 2012), but, to the best of our knowledge, there is not yet published research deriving from this corpus.

As far as Spanish is concerned, the interest on academic discourse has not a long tradition. The pioneering project was ADIEU (Vázquez, 2001), more focused on Spanish as a second language and including a collection of transcribed texts of oral presentations and master classes. The main interest has been in the differences between academic genres (Regueiro Rodríguez & Sáez Rivera, 2013; Sanz Álava, 2007; Perea Siller, 2013). In the studies on academic genres, the research around the School of Valparaíso stands out (Parodi, 2010). This team has compiled an academic corpus PUCV-2006 (http://www.elgrial.cl/) gathering texts form the academic and professional areas of four domains: industrial chemistry, construction, engineering, social work, and psychology. However, this corpus has not yet been used for the research of lexical phenomena. In the same vein, the reference handbook on academic and professional writing in Spanish, edited by Montolío Durán (2014), does not include any chapter entirely devoted to phraseology. The only previous work on academic lexical combinations in Spanish comes mainly from researchers who conducted contrastive studies in Spanish and English; that is, research focusing on non native speakers of English and dealing with the differences between academic lexical combinations in these two languages (see, among others, Tracy-Ventura et al., 2007; Cortes, 2008; Perales-Escudero & Swales, 2011; Pérez-Llantada, 2014).

Even if English is gaining ground in Spanish universities, Spanish is still the most used language in academic texts by university students. However, to be a native speaker of a language does not guarantee to be academically competent in this language: there is no native speaker of academic language and, therefore, the competence in academic writing has to be learnt. Despite of writing in their native language, the academic writings of university students show often certain deficiencies, many of which come from a poor knowledge of ALC. If the difficulty is considerable for students who write in their L1, the challenge is still bigger in an L2. The growing number of foreigner students in the Spanish universities has shown the need of lexical resources which help them in their academic writing. Furthermore, these resources could also serve to improve the academic writing of experts researchers, since, due to internationalization, their academic L1 begins to be damaged (Johansson Kokkinakis et al., 2012).

The research presented in this paper forms part of a project that intends to fill that gap: we aim to build a combined dictionary-corpus tool in accordance with the current trends in lexicography, where resources provide lexical information in the form of a concordance program exploiting language corpora, instead of doing so only in the form of a dictionary (Asmussen, 2013; Paquot, 2012). Our focus is the discourse and phraseological conventions of academic Spanish in different domains, as we will explain later. In order to build up a useful resource, we need also to study the academic writing of students and examine the differences between the command of novice and expert writing with respect to ALC. The research questions behind the whole project are very similar to those presented by Cortes (2004):

1. Which are the most frequent ALC in published academic writing?
2. How are these ALC classified in phraseological and functional terms? Are there more collocations, idioms or formulas? Can the functional classifications thought for English lexical bundles by Biber et al. (1999) or Hyland (2008) serve for Spanish ALC?
3. Are there any significant differences of these ALC between disciplines?
4. Are the ALC used by university students? Are there differences in Bachelor's degree and Master's degree students? In different disciplines?

To answer these questions we simultaneously take two perspectives: Corpus Linguistics and Phraseology. Corpus Linguistics provides us with tools (frequency and other measures of lexical association) useful to identify ALC candidates. Phraseology allows for selecting among these candidates by applying some criteria issued mainly from the Meaning-Text Theory (MTT) (Mel'čuk, 2015), keeping in mind that the final aim is to build up a useful tool for writing academic texts in Spanish.

This paper is structured as follows. Section 2 focuses on the different types of ALCs and tries to establish distinctions among the messy characterization of phraseological expressions present in the literature. Section 3 provides a description of the methodology we are using, along with a presentation of the expert academic corpus that we are studying and of the compilation of the student corpus. Section 4 is devoted to the description of the tool's design. There, we present how the corpus and the lexical database are intertwined and we provide some samples of prototype entries for different kinds of ALC. Finally, in Section 5, we draw some conclusions on the presented work and give future lines of research.

2. Academic phraseology: defining ALC

It is well known that there is not an established terminology to distinguish between different multiword units. Depending on different linguistic schools or traditions, what is a collocation for an author is a free phrase for another (e.g. *the results suggest*) and what is an idiom (*locución* in Spanish) from one perspective is considered a discourse marker from another (e.g. *in conclusion*), which is not contradictory. It is not only an issue of using different terms for the same concept, but also of labeling different concepts by means of the same term. The disagreement on the taxonomies of multiword units is not specific of research in the academic genre, but is common in phraseological inquiries, regardless of textual type.

In order to determine the phraseological nature of multiword sequences and to adscribe them to a phraseological category, we will adopt the tenets of Meaning-Text theory (Mel'čuk, 2015). Within this theoretical framework, two criteria are of paramount importance to ascertain whether a certain lexical combination is phraseological: its *compositionality* (not to be confused with its *transparency*) and the free or conditioned choice of its components. Compositionality is a property whereby the meaning of a given expression is the result of adding up the meanings of its constituent parts. Compositionality, which is production-oriented, should not be confused with transparency, which has to do with the understandabilty of an expression. Thus, an expression that is fully transparent is necessarily compositional, but the inverse is not true; for example, if a speaker does not know what the verb *respectar* means, he cannot guess the meaning of the expression *en*

lo que respecta a 'concerning X', even if this expression is fully compositional. Therefore, a compositional expression can be non-transparent.

If an expression is fully compositional, it could still be considered phraseological, as long as its components are not *freely* chosen or combined. When a phrase is *free*, each of its lexical components is selected strictly due to its meaning, independently of the lexical identity of other components (Mel'čuk, 2012, 33). The adjective *free* must be then understood strictly as allowing the selection of one lexical unit independently of the other lexical components of the same expression (Mel'čuk, 2012, 33). Thus, in the Spanish phrases *la probabilidad de que* ('the probability that') or *al revisar la selección* ('when reviewing the selection'), each of their lexical components is selected because of its meaning and combinatorial properties in conformity with the corresponding rules of Spanish (Mel'čuk, 2015, 59).

In contrast, a non-free phrase (*lexical phraseme*, in MTT terminology) is not constructed out of its lexical components by selecting each individually and arranging them according to the standard rules of L. Other non-standard rules specify a non-free phrase as a whole. The constraints that operate in the production of a non-free phrase can take place at different levels. Depending on compositionality and the type of constraint, our theoretical framework distinguishes several types of lexical phrasemes. The following pages focus on three types: idioms, collocations and formulas, which are the ALCs that our lexical tool will include. In what follows we are going to present each in turn.

2.1 ALC: idioms

We consider an idiom any non-free phrase if it is non compositional. An idiom is selected as a whole: from its semantic representation, a special rule maps its meaning to a single lexical node in a syntactic representation. Thus, for example, *en conclusión* (or its English equivalent, *in conclusion*) is one lexical unit, yet made up of two words. It should be the headword of its own lexicographical entry with its definition, its part of speech, and all relevant combinatorial information.

Idioms are very frequent in academic prose, especially those considered discourse markers from other perspectives: *en consecuencia* ('consequently'), *al contrario* ('on the contrary'), *por otra parte* ('on the other hand'), etc., although we encounter other types, such as verbal idioms, like *llevar a cabo* ('carry out'), *dar lugar* ('bring about') or *tener en cuenta* ('take into account'), nominal idioms such as *punto de vista* ('point of view') and — fewer — adjectival idioms.

There is also an overlap between idioms and lexical bundles; for instance, *en relación con* ('in relation with') is traditionally included in Spanish dictionaries as a prepositional idiom.

2.2 ALC: collocations

Unlike idioms, collocations are compositional. They are composed of two lexical units: the *base*, the selection of which is semantically-driven and the *collocate*, which is chosen not only on semantical, but also on lexical grounds (Mel'čuk, 1996, 37). Thus, in the verbal collocation *sacar conclusiones* ('draw conclusions'), the base *conclusión* conditions

the choice of the collocate *sacar* (lit. 'pull out'). If the base were *decisión* ('decision'), the choice of the support verb would be different: namely, *tomar* (lit. 'take'). Even if collocations are compositional, they are phraseological because the choice of one of its components is constrained by the other. The lexicographical description of each collocation should be made under the entry of the base. We intend that the user of our lexical tool will be able to recover information on collocates by means of an inverse search (see Section 4).

In academic prose, we focus on verbal collocations with the syntactic pattern verb-object, and also subject-verb, as *problema* and *estribar* in e.g. *el problema estriba* ('the problem lies'). Adjectival collocations are also object of our interest: *conclusión correcta*, *obvia, lógica*, *contraria*.

2.3 ALC: formulas

Formulas (*formulemes* in terms of MTT) are also compositional: *en otras palabras* ('in other words'), *es bien conocido que* ('it is well known that'), *no hay que olvidar que* ('we should not forget that'), *como se ha señalado previamente* ('as previously stated'), etc. However, both the meaning of a formula and its lexical implementation are constrained. Mel'čuk (2015) points out that if a speaker has the intention: 'I will now express the same content I have just expressed, but using different words', he cannot select the meaning 'I signal that the following fragment of my speech means the same as the preceding fragment' (the meaning of expressions such as *in other words* or *to put it differently* has more to do with the notion of 'rephrasing' than with that of 'repeating ideas'). From the former meaning, the speaker is not free to select any fairly synonymous expression, such as *using some different expressions* or *I say this in a different way*, because these expressions are not natural in English. The same happens in Spanish. From the same semantic representation, a Spanish speaker could produce *en otras palabras* and *dicho de otro modo/otra forma/otra manera*, but not *por ponerlo diferente* (cf. Eng. 'to put it differently').

As shown, formulas are doubly constrained. However, they do not need a lexicographic definition because a formula means exactly what it says. They need, in contrast, a description of its discourse function, especially in academic discourse (Cortes, 2004, 241). Thus, users of a lexical tool as the one proposed could obtain, for instance, different ways to emphasize a statement; e.g., *hay que destacar* ('it is necessary to stand out'), *es importante subrayar* ('it is important to emphasize'), *mención especial merece* ('it is worth mentioning'), etc.

Even though academic texts swarm with formulas, their theoretical status is not sufficiently clear. English dictionaries collect formulas such as *in other words*, (*and*) *what's more*, etc., but the Spanish dictionaries do not. For example, *en lo que respecta* ('in what concerns') appears under the headword *respectar* but this verb is defective and is used only in this expression with the variant (*por lo que respecta*). Other formulas are perceived as having less "lexical entity". Thus, recurrent sequences of academic discourse such as Engl. *the aim of this work is*, *the results suggest*, *this study has shown that*, among others, are not collected as phrases in any academic English dictionary, although they appear in lists of lexical bundles.

This third category is perhaps the one having more in common with the concept of *lexical bundle*, which has gained increasing acceptance in current research in academic

discourse. However, in contrast to our formulas, lexical bundles are not defined on account of the choice or their components and their compositionality, but on purely distributional terms: *lexical bundles* are contiguous word sequences or n-grams that display a minimum frequency (usually from 10 to 40 occurrences per million words) and a minimum dispersion in corpora (cf. Biber et al., 1999). Apart from the theoretical differences, it could be relatively safely stated that all formulas are lexical bundles, but the opposite is not always the case. For example, *la probabilidad de que* ('the probability that') can be considered as a lexical bundle by virtue of its recurrence and dispersion, but from our perspective this multiword sequence is not phraseologically relevant. As we will explain in the next section, the techniques developed to identify and extract lexical bundles are useful for our research, but lexical bundles themselves are, so to speak, raw materials that have to be processed before being included in our lexical tool.

2.4 Recapitulation

The limits between the three different ALCs are not always completely clear. The compositionality draws a boundary between idioms, on the one hand, and collocations and formulas, on the other. When one of the components is a grammatical word, the distinction is less obvious. For instance, *sin duda* ('without a doubt') seems to be compositional because its meaning includes 'without' and 'doubt'. However, its meaning includes also a discourse semantic component that emphasizes speaker's statements.

ALCs can merge sometimes. This happens, for instance, when a formula contains a collocation. In academic prose it is frequent to encounter formulas such as *la pregunta que nos tenemos que formular* ('the question we should ask'), which includes the verbal collocation *formular una pregunta* ('to ask a question').

In our lexical tool, all formulas and some idioms will receive a discourse function. Collocations will be included in the entries of their respective bases and will not be associated to any specific discourse function, since arguably those are associated to specific sequences of words. E.g., the lexical entry for the base *pregunta* will include all its collocates, but its collocations will not have discourse function because this one is associated only to a concrete sequences of words.

3. A not so radical corpus-driven approach to academic phraseology

Our methodology is corpus-driven, but not as radical as the one adopted by Biber (2009, 281). Even if corpora play a leading role in the construction of our lexical tool, we need to filter out corpus output with phraseological criteria, which makes human intervention necessary. This section describes the corpora used for our study and the methodology applied to extract information from them.

3.1 Corpus description

We need two types of corpora: first, an expert academic corpus in order to obtain the list of ALCs for our lexical tool. The corpus used is the Spanish part of the Spanish–English Research Article Corpus (SERAC 2.0), a 5.7-million word compilation of 1,056

research articles (RAs). It includes 360 L1 RAs in Spanish published by Spanish scholars in peer-reviewed journals targeted at a national-based scholarly readership (Pérez-Llantada, 2014). The corpus contains about two million running words. It is divided into four sections, namely: Arts and Humanities, Biological and Health Sciences, Physical Sciences and Engineering, Social Sciences and Education.

Second, we have begun to compile a novice academic corpus for Spanish with a view to building a resource similar to BAWE (Gardner & Nesi, 2013) or MICUSP (Römer & O'Donnell, in preparation) for English. We are compiling Bachelor's and Master's degree theses of Spanish university students in the same areas as the expert corpus. The identification of student's difficulties with ALC in this corpus will be key for the design of the lexical tool that we project.

3.2 Quantitative approach

Currently, we have completed the compilation of a list of academic Spanish words and the extraction of academic collocations, formulas and idioms is in progress.

The Spanish Academic Word List (SAWL) consists of about 1,000 lemmas of content words (nouns, verbs, adjectives and adverbs) and has been extracted following two criteria (similar to Coxhead (2000) or Paquot (2010), among others): (a) the *keyness* of the forms extracted and (b) their dispersion. The keyness of the lemmas has been determined by comparing their distribution in the expert corpus and in a contrast corpus (the narrative part of the LEXESP corpus, Sebastián-Gallés et al., 2000) by means of the Wilcoxon-Mann-Whitney test (cf. Kilgarriff, 2001; Lijffijt et al., 2014). We retained those items with a significance of $p < 0.001$. To avoid vocabulary specific of only a certain thematic field, we have controlled for dispersion using Gries's DP coefficient (Gries, 2008) by including only those items with a value of 0.5 or less (cf. Durrant, 2014).

This vocabulary list will be further manually filtered assessing the collocational and the discourse productivity of its items: if a word of the SAWL is productive as a basis of many collocations and it is a member of formulas with discourse functions, it will be candidate to be part of the macrostructure of the lexical database. Collocations will be extracted by using dependency parsing and measures of lexical association. Such extraction procedure in all probability will yield combinations with different phraseological status (e.g. collocations such as *extraer conclusiones* and idioms such as *tener en cuenta*) that will have to be manually sorted out.

The extraction of recurring n-grams seems a strategy more suitable to extract formulas and certain types of idioms such as prepositional or adverb phrases, made up of contiguous word sequences (e.g., *a través de, sin embargo, no obstante*; cf. Tutin & Kraif, 2017). A preliminary analysis has put into question the suitability of keyness when filtering lists of n-grams for our current purposes: such filtering yields poor recall values, since a considerable amount of phraseologically interesting multiword chains do not reach significance thresholds. Likewise, while frequency thresholds conventionally used for retrieving lexical bundles produce acceptable results with 4-grams, additional measures seem to be necessary in order to get rid of 2-grams and 3-grams of dubious interest (backwards transition probability as proposed in Appel & Trofimovich (2015), seem to get the best results in our n-gram list).

3.3 Filtering and enriching raw data

We adopt a mixed-method approach similar to Simpson-Vlach & Ellis (2010) or Ackermann & Chen (2013), who also combined statistical information and human judgment when compiling their respective lists of academic lexical combinations for English. After obtaining n-gram lists by using statistical measures, we will apply phraseological criteria to discriminate between idioms, collocations and formulas. The classification is necessary because each type of ALC requires a different lexicographic description, as we will show in Section 4. Only idioms and formulas are enriched with discourse functions.

The typology of discourse functions is being obtained following a bottom-up approach. Even if we start from previous classifications of lexical bundles in English (Hyland, 2008; Simpson-Vlach & Ellis, 2010; Salazar, 2014), their taxonomy cannot be directly imported to Spanish ALCs. Most of them distinguish between three main functions: 1) describing research, 2) organizing text and 3) conveying the author's stance and interacting with reader (Salazar et al., 2013, 45). Each function has a long list of subfunctions that are not always easily interpretable for a potential user of a lexical tool. For instance, the function "framing", used by Hyland (2008) or Salazar (2014), groups together lexical bundles such as *with respect to*, *with the exception of*. The function framing serves to "situate arguments by specifying limiting conditions" (Salazar, 2014, 52). Even if the cited bundles fit within this definition, it might be useful to provide the user with more specific information about when to use each one. A similar objection can be raised against putting together *it should be noted that*, *see Figure 1, as seen in* under the function "address readers directly" (Salazar, 2014, 52). These formulas do indeed address readers directly, but they do not have the same discourse function in an academic text: the first one boosts the statement that follows, whereas the other two point out specific fragments of the text.

We aim to build a typology with the main discourse functions in academic writing more oriented to the user, following Gilquin et al. (2007) and Prat Ferrer & Peña Delgado (2015) with simple and clear headings (e.g., "how to begin", "changing subject", "presenting conclusions", etc. see Figures 1–4). We adopted the following process: first a sample of articles included in the expert corpus has been examined in order to put forward a list of discourse functions. We are studying which formulas and idioms fulfill these functions by checking the contexts where they occur. It is likely that the typology of discourse functions devised after the qualitative revision of the mentioned sample will be improved during this process. The final product will be a database of ALCs associated with discourse functions, rather than a corpus annotated with discourse functions.

The assignment of discourse functions cannot be made automatically, save perhaps some exceptions. Thus, a formula such as *esta es la principal conclusion* ('this is the main conclusion') is not necessarily used to conclude. Its context must be examined in order to verify whether, for instance, it is mainly employed to introduce the conclusions of a paper or to refer to the research of other authors, as in *esta es la principal conclusión a la que llega el estudio X*, etc. We are aware that this manual assignment is slow, but we project to get complete products by working discourse functions. In this way, we can obtain finished descriptions in different phases of our project, such as "ALC which serve to conclude", "ALC which serve to emphasize", etc.

4. Design of lexical tool HARTA[1]

We aim to build a combined dictionary-corpus tool in accordance with the current trends in Lexicography, where resources can provide lexical information by means of concordances coming from corpora, ins addition of doing so only in the form of a dictionary (Asmussen, 2013; Verlinde & Peeters, 2012). The corpus is intertwined with the lexical database, because, in many cases, user queries are more easily answered by showing examples of a given ALC in corpus, rather than by offering a whole lexicographic description. In the last few years, several authors recommend to expose both L2 learners and novice writers to corpus-based evidence (Cortes, 2013; Pérez-Llantada, 2014; Cotos, 2014). More recently, Laso & John (2017) have taken a step beyond awareness-raising by investigating the influence of corpus consultation on the written production.

4.1 Macrostructure of HARTA

The macrostructure of HARTA is only partially based on the list SAWL. The selected headwords must fulfill a discourse function or be part of an ALC fulfilling one. There will be two kinds of lexicographical entries: proper entries for single and multiword lexical units, with all the information an entry is supposed to contain in an MTT framed dictionary (semantic, syntactic and combinatorial), and *ad hoc* entries for formulas. As explained above, many formulas are not properly a lexical entity, but it is useful for the user to access them through their discourse function. Thus, for instance, the noun *resultado* ('result') is chosen to be part of the macrostructure and will receive a whole entry because this noun is part of several formulas fulfilling discourse functions (*estos resultados sugieren/indican que*). Likewise, the idiom *punto de vista* ('point of view') will receive an entry because it is part of several formulas used to cite or to convey the author's perspective (*desde nuestro punto de vista*). Some idioms are used to serve a discourse function as a whole, such as *en conclusión* ('in conclusion') and, therefore, they will be provided with a proper entry also. For formulas we will choose a canonical form on criteria similar to those employed by Salazar (2014) to establish *prototypical bundles*.

4.2 Microstructure of HARTA

The whole entry includes information of two types: 1) the core information, consisting of semantic and combinatorial information about the lexical unit and 2) the usage information, including frequency, disciplines in which the unit occurs, etc., and access to the corpora (see Figure 1).

The entry for a formula contains the following fields (see Figure 2):

1. Discourse functions. A formula can have more than one function: e.g. as Salazar et al. (2013, 46) point out *these results suggest* serves to draw conclusions, but involves also the function of hedging due to the use of mitigating verb *suggest*.
2. Disciplines where the formula appears. Some disciplines are more prone than other to use a lofty style. Thus, a formula such a *mención especial merece* will probably be less frequent in Sciences than in Literature research.

[1] HARTA stands for *Herramienta de Ayuda a la Redacción de Textos académicos* ('tool of help for writing academic texts').

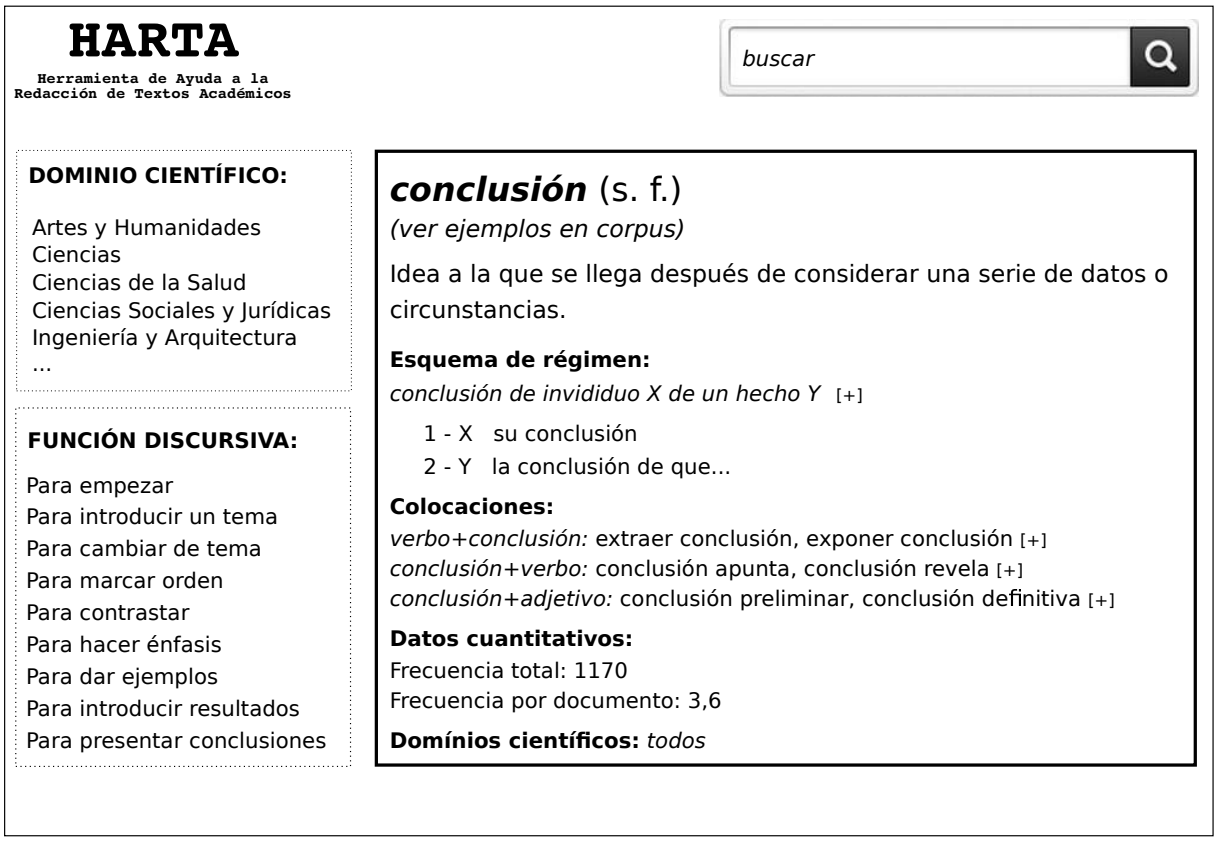

Figure 1: Entry for the noun *conclusión*

3. Frequency of co-occurrence. It is useful information for the user to know if the formula expressing a given discourse function is more or less productive than others.
4. The sections of the research article where the formula appears. We have marked up the sections of the text included in the expert corpus (abstract, introduction, body (method, result, discussion), conclusion). As Salazar et al. (2013, 49) pointed out, the discourse function can vary according to the section of the text. For instance, the formula *in accordance with* has the function of describing a procedure in the Methods section, whereas it is used to present the results from previous studies in the Discussion section.

Any lexical component of a formula will have a hyperlink to its own entry or trigger another kind of search. E.g., for the formula in Figure 2, there would be a hyperlink to the information associated to the idiom *tener en cuenta* ('to take into account').

4.3 Different access to the information

There will be two main search types: 1) the discourse function search and 2) the word search.

In the discourse function search, the user will be able to select a given function and get all the formulas fulfilling this function. In Figure 3 the user clicks on *para hacer énfasis* ('to emphasize'), and the tool provides a list of formulas (in their canonical form) which can be ordered alphabetically or by frequency. If the user clicks on each formula, he sees the entry (see Figure 2).

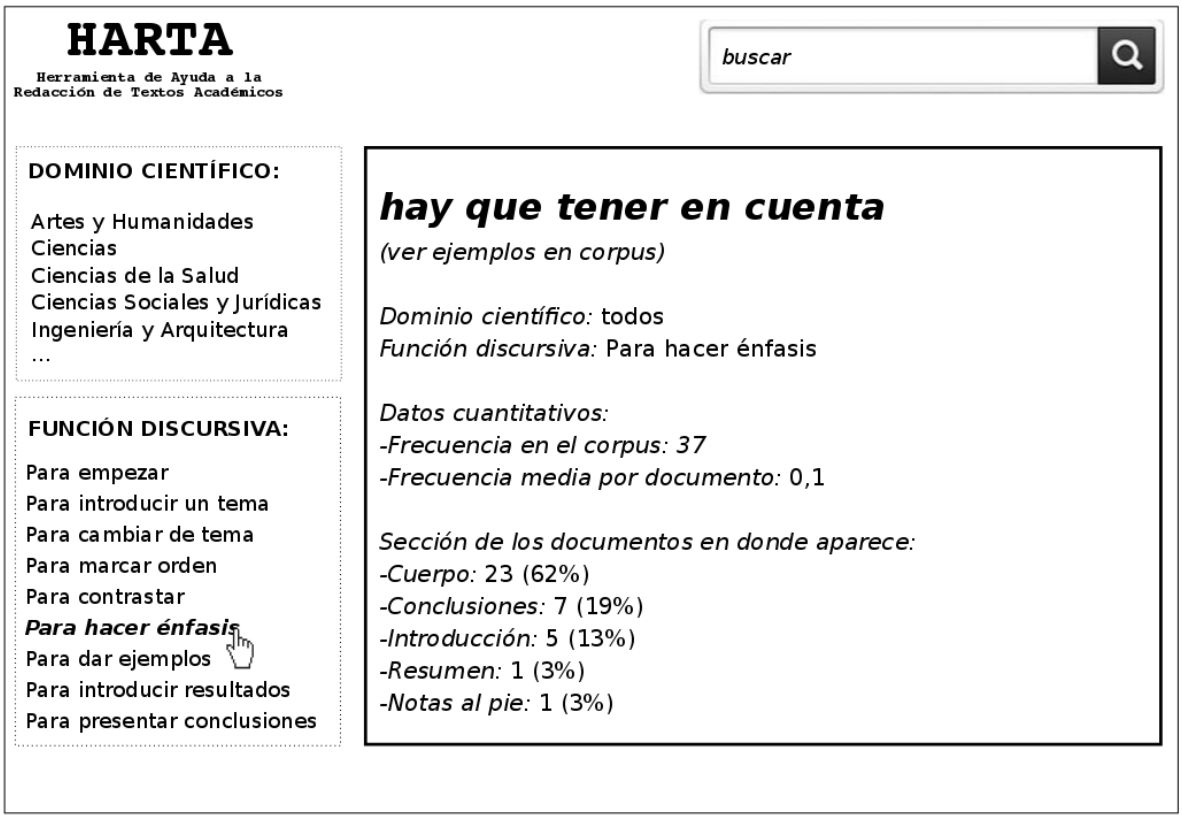

Figure 2: Entry for a formula

Furthermore, information will be accessible through word search (Figure 4). For example, if the user wants to obtain information on the noun *resultado* ('result'), the interface will provide access to its entry, if there is one, or to the formulas, idioms and collocations where it occurs. The entry for the noun *resultado* displays links to its collocations. More information will be found when clicking on the entry (see Figure 1 where you can see the proper entry for *conclusion*).

If the queried word has no proper entry, the interface will provide the formulas and the collocations in which it occurs. For instance, if an user looks up for the verb *sugerir* ('suggest'), the inferface would provide the formulas and all nouns which are the subject of this verb in collocations: *autor*, *análisis, dato, experimento*, *resultado*, etc. It should be noted that this information is what a search on a collocational database returns, not the static information included in an entry. In our theoretical framework we claim that collocational information must be described in the base's entry but should be recoverable both through the base and the collocate.[2]

5. Conclusions

This paper has presented an ongoing research on academic lexical combinations in Spanish with the aim of building a lexical resource accessible on the web. In contrast to other

[2] This is the policy that we use in the compiling of the Spanish collocation dictionary DiCE (http://www.dicesp.com/). We will build entries for bases, but information for collocates will be recoverable through special searches (Alonso-Ramos, 2016; Alonso-Ramos et al., 2010).

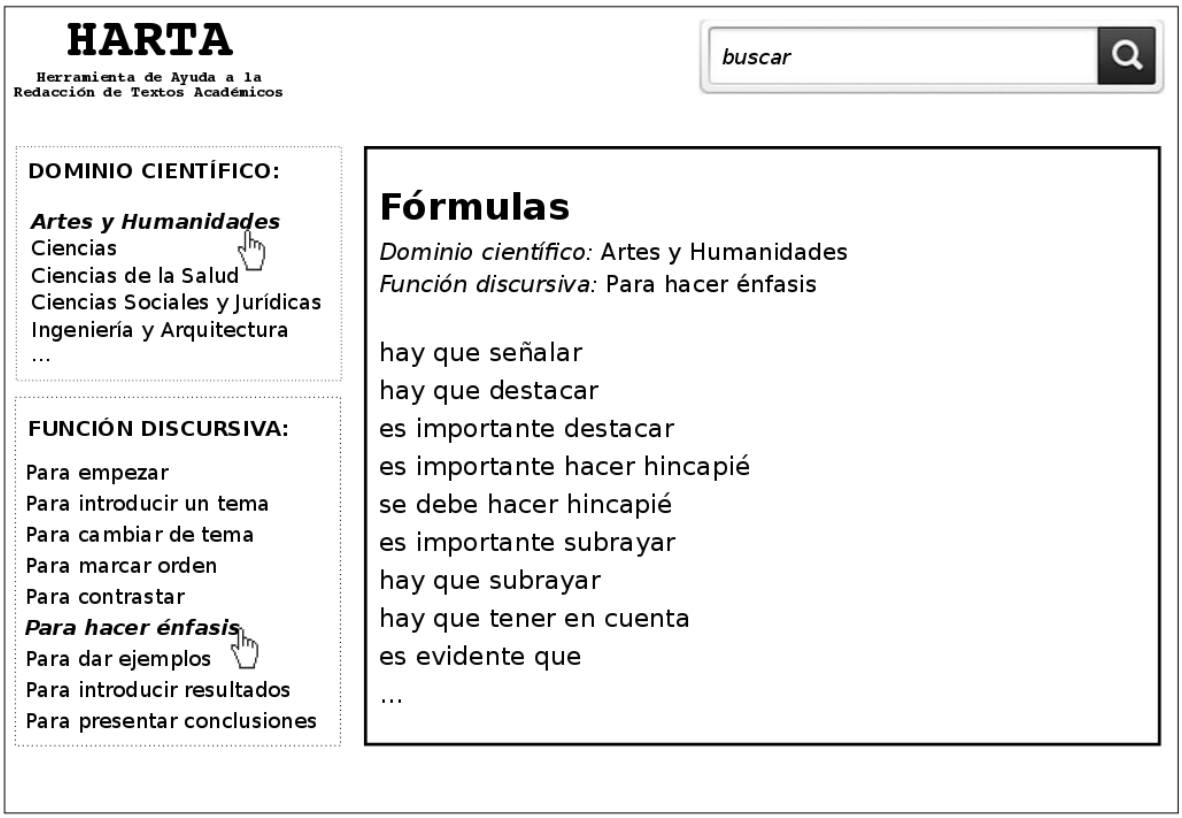

Figure 3: Discourse function search

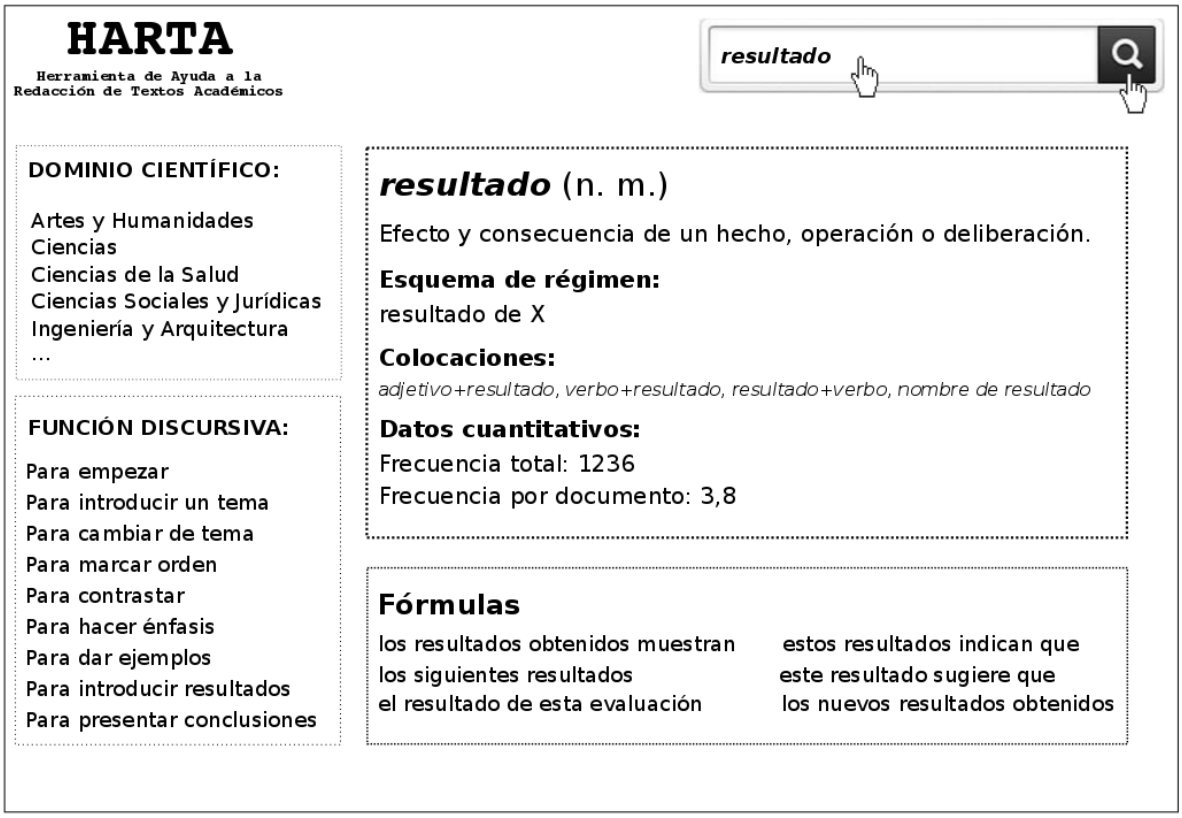

Figure 4: Word search

similar tools, such as LEAD (Paquot, 2012) or ScieLex (Verdaguer & Salazar, 2013), we intend a finer classification of phraseological units, since we rely on a theoretical framework that provides the necessary theoretical tools for the endeavour. We are aware that such distinctions involve a longer process. However, we project to get a product of increasing completeness along the successive stages of our research by devising an exhaustive classification of discourse functions. We believe that the final user will appreciate more a rich entry than lists of lexical bundles organized by mere frequency. In the meantime, access to the expert corpus will be profitable for any user.

We will better adapt to user needs when we have analyzed the student corpus. Differences in frequency of use between expert and novice writers will provide clues as to the difficulties faced by the latter and, accordingly, the type of information that should be given priority in the different entries. This analysis can also provide teaching material devoted to novice writers such as Salazar (2014) proposes.

6. Acknowledgements

The work presented in this paper has been partially supported by the Spanish Ministry of Economy and Competitiveness (MINECO), by the FEDER Funds of the European Commission under the contract number FFI2016-78299-P, by a postdoctoral fellowship granted by the Galician Government (POS-A/2013/191), and by a *Juan de la Cierva formación* grant (FJCI-2014-22853).

7. 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, pp. 235–247.

Alonso-Ramos, M. (2016). Learning resources for Spanish collocations: From a dictionary towards a writing assistant. In B.S. Vilas (ed.) *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching, volume C of Mémoires de la Société Néophilologique de Helsinki*. Helsinki, Finland: Société Néophilologique de Helsinki, pp. 65–95.

Alonso-Ramos, M., Nishikawa, A. & Vincze, O. (2010). DiCE in the web: An online Spanish collocation dictionary. In S. Granger & M. Paquot (eds.) *eLexicograpy in the 21st century: New Challenges, New Applications. Proceedings of eLex 2009*. Louvain-la-Neuve: Presses universitaires de Louvain, pp. 367–368.

Appel, R. & Trofimovich, P. (2015). Transitional probability predicts native and non-native use of formulaic sequences. *International Journal of Applied Linguistics*, 27, pp. 24–43.

Asmussen, J. (2013). Combined Products: Dictionary and Corpus. In R. Gouws, U. Heid, W. Sheweickard & H. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin/Boston: De Gruyter Mouton, pp. 1081–90.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3), pp. 275–311.

Biber, D., Conrad, S. & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), pp. 371–405.

Biber, D., Johansson, S., Leech, G., Conrad, S. & E., F. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson.

Cavalla, C. & Loiseau, P. (2014). Scientext comme corpus pour l'enseignement. In Tutin & Grosman (eds.) *L'écrit Scientifique: Du Lexique Au Discours*. Rennes: Presse universitaire de Rennes, pp. 163–180.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4), pp. 397–423.

Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora*, 3, pp. 43–58.

Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes*, 12, pp. 33–43.

Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL*, 26, pp. 202–224.

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

Durrant, P. (2014). Discipline and level specificity in university students' Written vocabulary. *Applied Linguistics*, 35(3), pp. 328–356.

Gardner, S. & Nesi, H. (2013). A classification of genre families in university student writing. *Applied Linguistics*, 34(1), pp. 1–29.

Gilquin, G., Granger, S. & Paquot, M. (2007). Writing sections. In M. Rundell (ed.) *Macmillan English dictionary for advanced learners*. Oxford: Macmillan Education, 2 edition, pp. 1–29.

Gries, S. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), pp. 403–437.

Hyland, K. (2008). As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27, pp. 4–21.

Johansson Kokkinakis, S., Sköldberg, E., Henriksen, B., Kinn, K. & Bondi Johannessen, J. (2012). Developing Academic Word Lists for Swedish, Norwegian and Danish – a Joint Research Project. In R.V. Fjeld & J.M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress*. Oslo: Department of Linguistics and Scandinavian Studies. University of Oslo, pp. 563–569.

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1), pp. 97–133.

Krause Kilian, C. & Dias Loguercio, S. (2015). Fraseologias de gênero em resumos científicos de Linguística, Engenharia de Materiais e Ciências Econômicas. *Tradterm*, 26, pp. 241–267.

Laso, N. & John, S. (2017). The pedagogical benefits of a lexical database (SciE-Lex) to assist the production of publishable biomedical texts by EAL writers. *Ibérica*, 33, pp. 147–172.

Lea, D., Bull, V. & Webb, S. (eds.) (2014). *OLDAE: Oxford Learner's Dictionary of Academic English*. Oxford: Oxford University Press.

Lijffijt, J., Nevalainen, T., Saily, T., Papapetrou, P., Puolamaki, K. & Mannila, H. (2014). Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, 31(2), pp. 374–397.

McCarthy, M. & O'Dell, F. (2008). *Academic Vocabulary in Use: 50 Units of Academic Vocabulary Reference and Practice ; Self-study and Classroom Use*. Cambridge: Cambridge University Press.

Mel'čuk, I. (1996). Lexical functions: a tool for the description of lexical relations in the lexicon. In L. Wanner (ed.) *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam: John Benjamins, pp. 37–102.

Mel'čuk, I. (2012). Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology*, 3(1), pp. 31–56.

Mel'čuk, I. (2015). Clichés, an Understudied Subclass of Phrasemes. *Yearbook of Phraseology*, 5, pp. 35–50.

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

Paquot, M. (2010). *Academic vocabulary in learner writing: from extraction to analysis*. London/New York: Continuum.

Paquot, M. (2012). The LEAD dictionary-cum-writing aid: an integrated dictionary and corpus tool. In S. Granger & M. Paquot (eds.) *Electronic Lexicography*. Oxford University Press, pp. 163–185.

Parodi, G. (ed.) (2010). *Academic and Professional Discourse Genres in Spanish*. Amsterdam/Philadelphia: John Benjamins.

Perales-Escudero, M. & Swales, J. (2011). Tracing convergence and divergence in pairs of Spanish and English research article abstracts: The case of Ibérica. *Ibérica*, 21, pp. 49–70.

Perea Siller, F.c. (2013). *Comunicar en la Universidad. Descripción y metodología de los géneros académicos*. Córdoba: Universidad.

Pérez-Llantada, C. (2014). Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes*, 14, pp. 84–94.

Prat Ferrer, J. & Peña Delgado, A. (2015). *Manual de escritura académica*. Madrid: Ediciones Paraninfo.

Regueiro Rodríguez, M. & Sáez Rivera, D. (2013). *El Español Académico. Guía Práctica Para La Elaboración de Textos Académicos*. Madrid: Arco Libros.

Römer, U. & O'Donnell, M.B. (in preparation). *MICUSP: A Corpus Resource for Exploring Proficient Student Writing across Disciplines*. Amsterdam: John Benjamins.

Salazar, D. (2014). *Lexical Bundles in Native and Non-native scientific writing*. Amsterdam/Philadelphia: John Benjamins.

Salazar, D., Verdaguer, I., Laso, N., Comelles, E., Castano, E. & Hilferty, J. (2013). Formal and functional variation of lexical bundles in biomedical English. In J.L. Isabel Verdaguer N. & D. Salazar (eds.) *Biomedical English: A corpus-based approach*, volume 56 of *Studies in Corpus Linguistics*. John Benjamins Publishing Company, pp. 39–54.

Sanz Álava, I. (2007). *El Español Profesional y Académico en el aula universitaria. El discurso oral y escrito*. Valencia: Tirant Lo Blanch.

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

Simpson-Vlach, R. & Ellis, N. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31(4), pp. 487–512.

Swales, J. & Feak, C. (2012). *Academic Writing for Graduate Students: Essential Tasks and Skills*. Michigan series in English for academic & professional purposes. Ann Arbor: University of Michigan Press.

Tracy-Ventura, N., Cortes, V. & Biber, D. (2007). Lexical bundles in speech and writing. In G. Parodi (ed.) *Working with Spanish Corpora*. London: Continuum, pp. 217–231.

Tutin, A. (2010). Showing phraseology in context: Onomasiological access to lexicogrammatical patterns in corpora of French scientific writings. In S. Granger & M. Paquot (eds.) *eLexicography in 21st century. New Challenges, new applications*. Louvain-la-neuve: Presses universitaires de Louvain, pp. 313–324.

Tutin, A. (2014). La phraséologie transdisciplinaire des écrits scientifiques: des collocations aux routines sémantico-rhétoriques. In A. Tutin & F. Grossman (eds.) *L'écrit scientifique: du lexique au discours. Autour de Scientext*. Rennes: PUR, pp. 24–44.

Tutin, A. & Grossmann, F. (2014). *L'écrit scientifique: du lexique au discours. Autour de Scientext*. Rennes: Presse universitaire de Rennes.

Tutin, A. & Kraif, O. (2017). Comparing Recurring Lexico-Syntactic Trees (RLTs) and Ngram Techniques for Extended Phraseology Extraction: a Corpus-based Study on French Scientific Articles. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) at the European Chapter of the Association for Computational Linguistics Conference (EACL 2017)*. Valencia: Association for Computational Linguistics, pp. 176–180.

Vázquez, G. (2001). *Guía didáctica del discurso académico escrito: ¿cómo se escribe una monografía?* Madrid: Edinumen.

Verdaguer, Laso, N. & Salazar, D. (eds.) (2013). *Biomedical English. A corpus-based approach*. Amsterdam/Philadelphia: John Benjamins.

Verlinde, S. & Peeters, G. (2012). Data access revisited: the Interactive Language Toolbox. In S. Granger & M. Paquot (eds.) *Electronic lexicography*. Oxford: Oxford University Press, pp. 147–162.