Bellaterra Journal of Teaching & Learning Language & Literature
Vol. 12(1), Feb-Mar 2019, 65-90

Lexical bundles in learner and expert academic writing

Noelia Navarro Gil
Universidad Complutense de Madrid and Universitat Internacional de Catalunya

Elena Martínez Caro
Universidad Complutense de Madrid

(Text received March 19 2019; accepted March 22 2019; final version March 28 2019)
DOI: https://doi.org/0.5565/rev/jtl3.794

Abstract: Lexical bundles (LBs) have been described as the 'building blocks of discourse'; in addition to being highly frequent in writing and reducing processing time for readers and writers, they also perform important functions in language. LB choice, however, can vary according to genre, discipline, and different sections of the same text, which poses a challenge for novice L2 writers. This paper explores the use of LBs in a learner corpus of bachelor dissertations written in English by Spanish L1 students in linguistics and medicine, and compares it with published research articles in the same disciplines. By focusing on the introduction and conclusion sections, we identify the most frequent 3-, 4- and 5-word bundles in the corpora, to later study their types, structures, and functions. The results show differences in the use of LBs across disciplines, genres and sections, suggesting pedagogical implications for the inclusion of LBs in the L2 writing curriculum.

Keywords: lexical bundles, research articles, academic writing, phraseology, learner corpus

Resumen: Los paquetes léxicos (LBs) se describen como 'bloques de construcción del discurso'. Además de ser muy frecuentes en el habla y la escritura, y de reducir el tiempo de procesamiento para lectores y escritores, también desempeñan funciones importantes en el lenguaje. Sin embargo, la elección de LBs puede variar según el género, la disciplina e incluso diferentes secciones del mismo texto, lo que plantea un desafío para escritores noveles. El presente artículo explora el uso de LBs en un corpus de trabajos de final de grado escritos en inglés por estudiantes españoles de lingüística y medicina. Este corpus se compara con artículos de investigación publicados en las mismas disciplinas. Centrándonos en las secciones de introducción y conclusión, identificamos los LBs de 3, 4 y 5 palabras más frecuentes en el corpus, para luego estudiar sus tipos, estructuras y funciones retóricas. Los resultados muestran diferencias en el uso de LBs entre disciplinas, géneros y secciones, lo que sugiere implicaciones pedagógicas para su inclusión en la enseñanza de la escritura académica en inglés.

Palabras clave: paquetes léxicos, artículos de investigación, escritura académica, fraseología, corpus de estudiantes

Resum: Els paquets lèxics (LBs) s'han descrit com a 'blocs de construcció del discurs'. A més de ser molt freqüents en l'escriptura, i de reduir el temps de processament per a lectors i escriptors, també exerceixen funcions importants en el llenguatge. No obstant això, l'elecció de LBs pot variar segons el gènere, la disciplina i

fins i tot diferents seccions del mateix text, plantejant un desafiament per a estudiants novells. El present article explora l'ús de LBs en un corpus de treballs de fi de grau escrits en anglès per estudiants espanyols de lingüística i medicina. Aquest corpus es compara amb articles de recerca publicats en les mateixes disciplines. Centrant-nos en les seccions d'introducció i conclusió, identifiquem els LBs de 3, 4 i 5 paraules més freqüents en el corpus, per a després estudiar els seus tipus, estructures i funcions retòriques. Els resultats mostren diferències en l'ús de LBs entre disciplines, gèneres i seccions, suggerint implicacions pedagògiques per a la seva inclusió en l'ensenyament de l'escriptura acadèmica en anglès.

Paraules clau: paquets lèxics, articles de recerca, escriptura acadèmica, fraseologia, corpus d'estudiants

Introduction

Over the last few decades, numerous corpus analyses have brought to the fore the fact that language is highly patterned (Hunston, 2002; Römer, 2010; Sinclair, 2005). Sequences such as *additional information* or *is one of the main*, especially common in particular registers, are 'ready to use' chunks, "stored and retrieved whole[s] from memory at the time of use" (Wray, 2002, p. 9) rather than generated item-by-item. These pre-fabricated units have been shown to facilitate production for authors and also save processing effort for readers and listeners (Nattinger & DeCarrico, 1992).

Collocations (see Ackermann & Chen, 2013; Nesselhauf, 2005), idioms (see Grant & Bauer, 2004), or lexical bundles, also called formulas, clusters, or chunks (see Biber et al., 1999; Cortes, 2004; Hyland, 2008), are some of the different subsets studied in phraseology (Granger & Paquot, 2008; Meunier & Granger, 2008).

Lexical bundles (henceforth LBs) were first identified by Biber and colleagues (Biber & Conrad 1999, Biber et al., 1999) and have been defined as "the most frequently recurring sequence of words" (Biber & Barbieri, 2007, p. 264), as well as "important building blocks of discourse" (p. 270). The identification of LBs in corpus studies has been primarily based on corpus-driven approaches of frequency and range, following the pioneering lexical bundle approach developed by Biber, Conrad, and Reppen (1999). In order to qualify as a lexical bundle, a sequence needs an occurrence of at least 20 or 40 times per million words (Biber & Barbieri, 2007; Chen & Baker, 2010; Cortes, 2004). Range of dispersion (i.e. the number of texts in which the bundle appears) is normally set at 3 or 5 texts or 10% of the texts in the corpus (Hyland, 2008). This criterion is used to guard "against idiosyncratic uses by individual speakers or authors" (Biber & Barbieri, 2007, p. 268).

Structurally, less than 5% of LBs represent complete structural units (Biber et al., 1999, p. 991), and are commonly used to bridge phrases (e.g. *in the case of*) or clauses (e.g. *I want to know*). Even though LBs are not structurally complete, they have been shown to perform major functions in discourse. They can also occupy different positions in a text. According to Hoey (2005, p. 13), lexical items "are primed to occur in or avoid certain positions within the discourse", which Hoey calls 'textual colligation', another feature that facilitates text processing and production.

Textual colligation analyses can help to reveal the interaction between positioning of LBs and discourse functions. In particular, there are two main sections of academic texts which tend to be highly conventional and contain certain LBs that help to accomplish rhetorical moves: these are the introduction and conclusion sections. Lexical items in these sections respond to genre and discipline conventions, since they reflect recurrent communicative purposes of a particular community. According to Bondi (2010, p. 99), "the ethos of the discipline –what the community considers appropriate methodology and relevant objectives– may have an impact on language choice". For example, finding the bundle *our study has shown* –which normally occurs in the conclusion section– earlier in the text (e.g. in the methods section) may strike the expert reader as an unusual practice and denote immaturity or foreignness on the part of the writer (Mur-Dueñas, 2011; Sheldon, 2018).

Each mode (e.g. written), genre (e.g. student essay), register (e.g. formal), and discipline (e.g. medicine) tends to "employ a distinct set of lexical bundles, associated with [its] typical communicative purposes" (Biber & Barbieri, 2007, p. 265). Thus, there seems to be no "single pool of lexical bundles" (p. 265) writers or speakers can draw on. In order to demonstrate membership in a given community, authors need to successfully use the LBs that are typical of that genre and discipline (Ädel & Erman, 2012). Writers who lack experience or exposure to the target language in a particular register may not choose the most appropriate expressions, and will not easily be accepted as an 'insider' of that community (Durrant & Mathews-Aydınlı, 2011; Hyland, 2008; Wray, 2002). Unfortunately, knowledge of phraseology does not seem to be something innate: it is indeed far from being a "language universal skill" (Pérez-Llantada, 2014, p. 85). Due to their quantity and diversity, L2 and novice writers may find LBs difficult to acquire and master (Liu, 2012); in this respect, problems such as underuse, overuse, or misuse (both structural and functional) of certain bundles have been reported in the literature (see Ädel & Erman 2012; Chen & Baker, 2010; Meunier & Granger, 2008).

The present study aims to further the understanding of phraseology in learner writing by exploring the use of LBs in the introduction and conclusion sections of bachelor dissertations (BDs) written in English by Spanish L1 university students in linguistics and medicine. In order to compare the frequency of form, structure, and function of these bundles, an expert corpus of research articles (RAs) in the same disciplines is used as the reference corpus. The comparisons will be made from both a quantitative point of view –applying a corpus-driven approach to identify bundles in the learner and the expert corpus– and a qualitative approach –classifying the bundles structurally and functionally in both corpora. This study hopes to contribute to the body of research that studies phraseology in academic writing, and to serve as a useful pedagogical resource for L2 learners of English who are trying to accommodate to the conventions of these specific disciplines.

Literature review

Among the numerous studies on LBs over the last decades, we find comparisons of different populations (e.g. native *vs.* non-native speakers or students *vs.* experts [Ädel & Erman, 2012; Appel & Wood, 2016; Chen & Baker, 2010; Durrant & Mathews-Aydınlı, 2011; Hyland, 2008]), genres (e.g. RAs *vs.* textbooks [Bondi, 2010; Römer, 2010]), disciplines (e.g. soft and hard sciences [Byrd & Coxhead, 2010; Cortes, 2004; Hyland, 2008; Liu, 2012]), registers (e.g. written *vs.* spoken [Biber & Barbieri, 2007]), languages (e.g. academic Spanish *vs.* academic English [Mur-Dueñas, 2011; Pérez-Llantada, 2014]), and different sections of a text (e.g. introduction and conclusion [Bondi, 2010; Sheldon, 2018]).

One recurrent finding is that English L2 writers' use of LBs does not always approximate the use by expert or native writers in terms of frequency, form, and function. For example, the masters and PhD candidates' writings explored in Hyland (2008) seemed to contain more impersonal clusters (i.e. avoiding stance), and more clusters in general compared to RA writers. The author suggests that less proficient writers rely on word combinations more often than expert writers. This finding contrasts with Durrant and Mathews-Aydınlı's (2011) study, in which student essays showed a lower production of formulas compared to RAs; differences regarding functional moves were also found. The authors suggest that the lack of attention paid to different genres and disciplines in academic writing education may account for these differences.

Another interesting finding in the literature in relation to our study is English L1 students' greater and more varied use of LBs, especially in structures such as unattended *this*,

existential *there*, hedging and negations, as compared to that of L2 university students, whose texts contained learner writing characteristic features, such as anticipatory *it* which, coupled with some informal lexical choices (e.g. *it is easy to*), pointed at register difficulties (see Ädel & Erman, 2012). In terms of functionality, L1 writers used stance more frequently than L2 writers. Interestingly, stance is one of the functions that differed the most among RA writers of the different languages (Spanish L1, English L2, and English L1) and disciplines studied in Pérez-Llantada (2014) and in Sheldon (2018): English L2 writers were found to transfer some of their L1 (Spanish) rhetorical practices into their L2 writing, which made their texts less interactional.

In order to investigate the use of LBs by Spanish L1 undergraduate learners writing in English in two different disciplines (i.e. linguistics and medicine) and sections (i.e. introduction and conclusion) in comparison with their expert-writer counterparts, three research questions were established in this study:

1. What are the most common lexical bundles in the introduction and conclusion sections of L2 learners' BDs in linguistics and medicine?
2. How are these lexical bundles used in terms of structure and function?
3. To what extent does the use of lexical bundles approximate or differ from published RAs in the same discipline?

Data collection

In order to carry out a quantitative and qualitative analysis of LBs in academic writing, two corpora were compiled: (1) a learner corpus of BDs in linguistics and medicine written in English by Spanish L1 undergraduates in their last year of studies, and (2) an expert corpus of RAs in the same disciplines published in English-medium and peer-reviewed academic journals[2]. The introduction and the conclusion sections of each text were extracted and saved as raw .txt files for their separate analysis. Table 1 describes the number of texts, tokens, types, and paragraphs per genre, discipline and section.

Table 1. The learner and the expert corpus

	BDs		RAs	
Discipline	Linguistics	Medicine	Linguistics	Medicine
Intro no. texts	10	10	25	25
Tokens	5,724	9,063	17,722	11,535
Types	1,409	2,367	3,057	2,717

Avg. words intro.	572.4	906.3	708.8	461.4
Avg. paragraphs	3.1	3.9	2.9	1.2
Concl. no. texts	10	10	25	25
Tokens	4,703	4,555	15,214	14,679
Types	1,370	1,353	2,771	3,005
Avg. words concl.	470.3	455.5	608.5	587.1
Avg. paragraphs	2.5	3.5	3.5	1.2
Total words	10,427	13,618	32,936	26,214

Extraction, filtering, and classification of lexical bundles

In the present study, a corpus-driven approach was adopted in order to retrieve LBs from the corpora –i.e. no previous assumptions were made with respect to the LBs' form or function, and no pre-defined list of bundles was used. The function 'cluster n-gram' in AntConc (Anthony, 2018) was used to extract LBs from the introduction and conclusion sections of the corpora. In terms of length, even though the 4-word scope is the most researched length in LB studies (Ädel & Erman, 2012), other studies suggest that many recurrent word combinations come in as 3-word bundles (Simpson-Vlach & Ellis, 2010); as a result, we decided to adopt a more inclusive approach and explore 3-, 4- and 5-word bundles in the texts. As for frequency, given the relatively small size of the corpora, the frequency cut-off was set at a minimum of 20 times per million words. In addition, a dispersion range of three texts, which represent three different writers, was set; the selection of these cut-off criteria was based on previous corpus studies (Ädel & Erman, 2012; Biber & Barbieri, 2007; Chen & Baker, 2010). It is important to note that when a bundle appears only on one of the lists, it does not mean that this specific bundle was not used at all by writers in the other subcorpora; as Ädel and Erman aptly put it, "it simply means that the frequency and dispersion criteria were not met in the other group's material" (2012, p. 85).

Once the LBs were automatically retrieved, manual filtering was required in order to eliminate undesired 'noise' that could affect the comparability of the multidisciplinary corpora –i.e. context-dependent bundles– and that could also inflate the results –i.e. overlapping bundles. To deal with the first type, context-dependent bundles such as *second language acquisition, native and non native speakers, stem cells management* were manually eliminated from the lists. The second type, overlapping bundles, involved combining sequences such as *as a result* and *as a result of,* in which *of* appears in brackets (e.g. *as a*

result (of)). Frequency, range, number of grams (i.e. number of words in the sequence), and section (introduction and/or conclusion) in which each bundle appeared were explored.

With regards to the grammatical structure of LBs, we initially followed Biber et al.'s (1999, pp. 1014-1024) classification, which distinguishes 12 structural categories for LBs in academic prose. After revising this and the taxonomy they provide for conversation, we present a taxonomy of 15 categories with four broad structural groups: 'noun phrase-based', 'prepositional phrase-based', 'verbal phrase-based', and 'other' bundles, following Chen and Baker (2010, p. 34), which can best integrate the LBs found in our data. The NP-based bundles include noun phrases, with or without post-modifier fragments (e.g. *the risk of*, *the most prevalent)*. PP-based bundles refer to those starting with a preposition plus a noun-phrase fragment (e.g. *of this paper*, *in addition to)*. The VP-based broad category is the largest group, integrating nine different structures, all containing a verb component –or an introducing element of a clause (e.g. *it is a*, *can be used to*, *to do so)*. Different structural patterns are included here, such as subject + predicator structures, other verb phrase combinations, such as those followed by a noun-phrase or prepositional-phrase fragment, those containing a passive verb, anticipatory *it* structures, and patterns with the clause-introducing elements *that* and *to*. This structural classification involved manual revision and classification of all bundles according to their structures (e.g. *the study of* was categorized as 'noun phrase with *of*-phrase fragment').

For the functional classification, on the other hand, we followed previous taxonomies (Biber, Conrad, & Cortes 2004; Cortes 2004; Hyland 2008) and classified all bundles into three main categories and their subcategories:

1) Research-oriented –also called referential in other models (e.g. Biber et al., 1999): LBs in this category help writers to situate, contextualize and describe their research. There are four main subcategories: 1) location (e.g. *at the beginning, at the university*), 2) procedure (e.g. *the use of the, the purpose of*), 3) quantification (e.g. *a part of, one of the most*), and 4) description (e.g. *the size of the, the nature of the*).

2) Text-oriented –also called discourse organizers (Biber et al., 1999): these LBs are concerned with the structure of the text and the interrelations established between the ideas presented. There are four main subcategories: 1) transitions (e.g. *on the other hand, in contrast to the*), 2) resultative (e.g. *as a result, due to the fact that*), 3) structuring (e.g. *in the next section, in this study*), and 4) framing (e.g. *with respect to, in the case of*).

3) Participant-oriented: LBs in this category show writers' attitudes towards the ideational content and address readers directly or indirectly. It comprises two main categories: 1) stance (e.g. *may be due to, are likely to*), and 2) engagement (e.g. *as can be seen, it should be noted*).

This functional classification was complex not only because the categorization involves subjectivity, but also because some LBs can perform more than one function (Liu, 2012). A concordance analysis was performed in order to see the extended context of certain bundles that seemed multifunctional. For example, *the basis of* is a 3-word bundle that can act as a research-oriented descriptive bundle, as in (1)

(1) Findings from such a study can form the basis of learner-relevant form-focused instruction. (LIN_RA01_I)[1]

But, when this sequence is part of the 4-word bundle *on the basis of,* it can mark a text-oriented resultative relationship, as in (2)

(2) Other linguistic accounts differentiate the two forms on the basis of information status, particularly in terms of topic. (LIN_RA15_I)

For those cases in which the authors could not agree on the categorization, even after analyzing their extended context, previous literature that included examples on LBs and their functional categories was consulted (Cortes, 2004; Hyland, 2008; Pérez-Llantada, 2014). These structural and functional classifications allowed us to better understand the use of LBs in the corpora studied.

Results and discussion

The results of the analysis of LBs are reported on as follows. First, the most frequent LBs in the introduction and conclusion sections of BDs and RAs in medicine and linguistics are explored. Convergent bundles (i.e. those bundles that appear on more than one list) are then presented. Finally, a second and more qualitative analysis of the structures and functions of bundles is presented, exploring the similarities and differences found in the corpora.

Frequency and convergence of lexical bundles in the corpus

There are a total of 218 different bundles in the corpus as a whole (for the full list, see Appendix 1) with a total frequency of 1,151 hits, which represents around 4.5% of the tokens in the corpus. The most frequent bundle is *the use of* with a raw frequency of 85 counts, which equals more than 1000 times per million words (pmw) in our corpus. Moreover, *the use of* appears in all genres and disciplines explored in this study, so it could be regarded as a core or convergent bundle, following Pérez-Llantada's (2014) nomenclature. It is noteworthy to mention that *the use of* appears in the conclusion section of the corpora 50 out of 85 times, clearly indicating a preference for the last sections of a text. RAs in linguistics (37) and in medicine (21) are the genres that contain more hits of *the use of*, very often paired with other nouns (*questions, tools, English, other alternatives, somatic stem cells*). This bundle seems to help writers to display results, as in (3) or limitations, as in (4).

(3) Trends for the social science fields indicate a reduction in the use of these informal features. (LIN_RA04_C)

(4) Another limitation was the use of asymptomatic microembolic signals as a surrogate marker. (MED_RA02_C)

The second most frequent bundle in the corpus is *in order to*, with a raw frequency of 62 counts, i.e. about 750 pmw. By contrast to *the use of*, this bundle appeared in the introduction sections of the texts more often, in particular, 39 out of 62 times. Taking into account the total number of words in each corpus, BDs in linguistics show a predominant use of this bundle (22 raw hits) followed by RAs in linguistics (24), BDs in medicine (12), and medical RAs (6). Different procedure verbs such as *address, determine, provide, show, solve, facilitate,* and *gain* are used after this bundle. *In order to* can help writers to emphasize the study's main objective or justification, as in (5) and (6) respectively.

(5) This study aims to analyse comprehension and production of false friends in students of English in a C1 level classroom in order to *explore* the influence of their mother tongue (L1) on a second language (L2). (LIN_BD10_I)

(6) Moreover, DTC's low prevalence requires the participation of a high number of medical centers in order to *obtain* a representative sample of patients. (MED_BD09_I)

The third most frequent bundle is yet another core bundle present in all subcorpora: *as well as* (43 hits). *As well as* appears more frequently in the introduction sections (24 times), and rather than just adding new information, this bundle helps writers to focalize and frame the ideas presented, as in (7) and (8):

(7) FN is a dimeric glycoprotein that is found in plasma as well as in the extracellular matrix (ECM) of various tissues (MED_RA03_I)

(8) Conclusions will be drawn to justify the analyzed usages of discursive strategies as well as the historical and social consequences that can derive from them. (LIN_BD02_I)

The use of, *in order to* and *as well as* are also included on Biber et al.'s (1999) list of the most common 3-word bundles in academic prose. These three bundles appear as well in the academic formulas list developed by Simpson-Vlach and Ellis (2010), and are in the top-200 'formulas worth teaching' (ranking 29, 4 and 5 respectively), which underlines their pedagogic relevance.

In terms of length, 3-word bundles were the most frequent in the corpus (85.7% of the total bundles), while 4- and, especially, 5-word bundles were scarcely used (10.2% and 3.9% respectively). This finding was similarly reported on in previous studies, such as Biber et al.'s (1999, p. 994), who found that 3-word bundles were much more frequent in academic prose (over 60,000 times pmw) than 4-word bundles (which occur over 5,000 pmw).

If we look at each subcorpus separately, in particular, we will find some interesting patterns. As can be seen in Table 2, BDs in medicine and linguistics have produced almost the same quantity of LBs in the introduction and conclusion sections (conclusions were a bit shorter in this genre compared to the introduction, which partially explains why they contain half the amount of LBs as introductions); this seems to point at a shared quantitative feature in the use of LBs between texts of two different disciplines but that belong to the same genre (BDs). This is only true, however, for the learner genre; RAs show a vastly different use of LBs in terms of frequency: even though there are the same number of texts, with similar

tokens for both introduction and conclusion sections, articles in linguistics contain almost three times more LBs than medical articles.

Table 2. Lexical bundles used in the learner and the expert corpus

	BDs		RAs	
Discipline	Linguistics	Medicine	Linguistics	Medicine
LBs intro.	23*	25	74	17
LBs concl.	10	8	73	23
Total LBs	33	33	147	40
Total freq.	156	131	674	190
N-grams	3-w (24)	3-w (30)	3-w (125)	3-w (38)
	4-w (5)	4-w (2)	4-w (17)	4-w (2)
	5-w (4)	5-w (1)	5-w (5)	5-w (0)

*all values are raw counts

This finding has been supported by previous literature on LBs in academic writing across disciplines (Hyland, 2008; Liu, 2012) and points towards a disciplinary difference: research suggests that soft-knowledge disciplines very often emphasize interpretative language in order to present persuasive arguments, compared to hard-knowledge disciplines, that tend to be more impersonal in their methods and discussions. The linguistic items that allow writers to achieve this objective are, more often than not, part of recurrent word combinations (e.g. *it is important to, has the potential to, it can be argued that, are likely to, seems to be, it should be, needs to be*), which can explain the prominent LB occurrences in linguistic RAs. Hyland (2008) reported that less mature writers had used LBs more often. This finding contrasts with our results, but only for one of the two disciplines: BDs in medicine do contain more LBs than RAs in the same discipline (3.3 *vs.* 1.6 bundles on average per text); particular characteristics of the BD genre with regards to its audience –for example, that of being an academic final assignment in which students need to show and convince their supervisors (as a superior entity) that they have acquired certain knowledge– can contrast with published RAs in which authors present information to peers (of more or less the same expertise) and could account for this quantitative difference.

Adopting another perspective, the comparison of all LBs lists has yielded an inventory of 35 shared bundles. Some of these bundles are shared in the introduction and conclusion section of the same subcorpus, but some are also shared between genres (BDs, RAs), disciplines (linguistics, medicine), and some of them appear in all lists, regardless of their

genre or discipline, what we call core bundles. These 35 bundles are the best candidates for general academic writing education and, supporting Pérez-Llantada (2014, p. 88), this inventory "might indicate that the writers have memorized these language sequences and routinized them in their writing practices". Table 3 shows convergent bundles in the corpora:

Table 3. Convergent LBs found in the corpora

	LIN BD intro	LIN RA intro	MED BD intro	MED RA intro
Discipline	in order to, in this paper, it has been, the fact that, the use of, there is a		as well as, in order to, the prevalence of, the risk of, the use of	
Genre	in order to, the use of	a number of, as well as, in order to, the use of, there is a	N/A	N/A
	LIN BD concl	LIN RA concl	MED BD concl	MED RA concl
Discipline	as well as, in order to, the fact that, the use of, this study has		the results of	
Genre	in order to, one of the	as well as, in this study, the current study, the present study, the use of, there is a	N/A	N/A
	LIN BD	LIN RA	MED BD	MED RA
Core LBs Intro/ concl.	in order to, the use of, as well as in order to, it has been, the fact that, the use of	as well as, based on the, differences in the, in order to, in relation to, in terms of, in this paper, in this study, of the most, some of the, the current study, the fact that, the importance of, the number of, the present study, the role of, the use of, there is a, understanding of the	in order to	a number of, as well as, the presence of, the prevalence of, the use of, there is a

As Table 3 shows, there are more LBs shared by discipline (linguistics shares 11 bundles, and medicine shares 6, in both introduction and conclusion sections) than by genre (BDs share 4 bundles, and RAs share 11). The fact that BDs, despite having noticeably fewer tokens than RAs, share more bundles with their respective discipline in a published genre than with their learner counterparts, indicates the important role disciplinary conventions play in academic writing.

If we look at specific bundles, as previously mentioned, *the use of* (85 hits), *in order to* (62) and *as well as* (43) are core bundles shared across all corpora in our study. Hyland (2008, p. 12) found a total of 5 core bundles across four disciplines (*on the other hand, as well as the, in the case of, at the same time,* and *the results of the*), which is somewhat similar to our results. In terms of bundles that appear in both the introduction and conclusion section of BDs and RAs, there are a total of 23 different bundles, 19 of which appear in the introduction and conclusion sections of RAs in linguistics; these items can be a useful resource for L2 writers of academic English. Convergent bundles not only vary in their grammatical structure but also in the discourse functions they perform, as we will see in the next section.

Structures and functions of lexical bundles in the corpus

Table 4 below shows the frequency of LBs per structure across genres and disciplines, taking the four broad groups and the 15 structural categories into consideration, and provides one illustrative example for each category. An important caveat to understand the discussion of the findings that follows is that the frequencies given refer to the *type* of bundles used and not to the number of times each bundle type was used (*raw frequency*).

Table 4. Frequency of LBs per structure: overall figures per genre and discipline (%)

LBs structures		BDs		RAs		
		LIN	MED	LIN	MED	Example
NP-based	Noun phrase with of-phrase	30.3	33.3	30.6	35.0	*the use of*
	Noun phrase with other post-modifier	9.1	3.0	6.8	5.0	*the fact that*
	Other noun phrase (fragment)	0	3.0	3.4	5.0	*the present study*
PP-based	Prepositional phrase with embedded of-	3.0	-	7.4	5.0	*in terms of*
	Other prep. phrase (fragment)	12.1	18.1	17	10	*of the most*
VP-based	Pronoun/noun phrase (fragment) + be	12.1	12.1	5.4	10	*there is a*
	Noun phrase (frag.) + verb phrase (except copula be)	12.1	3.0	2.0	2.5	*this study has*
	Verb phrase with active verb	-	3.0	2.0	-	*seems to be*
	Verb phrase + noun phrase fragment	-	9.1	2.0	-	*has the potential to*
	Verb phrase + prep. phrase fragment	-	-	2.7	-	*refer to the*
	Passive verb + prep. phrase fragment	6.0	3.0	3.4	12.5	*based on the*
	Anticipatory it + verb phrase/adjective phrase	-	-	2.7	-	*it can be argued that*
	(Verb phrase +) that-clause fragment	-	-	5.4	-	*that they are*

	(Verb/adjective +) to-(clause) fragment	9.1	6.0	6.8	10.0	*in order to*
Other	Other expressions	6.0	6.0	2.0	5.0	*as well as*
	Total	100	100	100	100	

As can be seen, there is a clear prevalence of NP-based bundles over the rest of structural categories in all corpora. This prevalence is especially evident in the expert corpus, in both linguistics and medicine (both with a total frequency of more than 40%), over the second most common group of structures, the VP-based bundles. The PP-based categories rank in the third position in all four subcorpora. It is worth looking at specific rather than general structural categories to obtain a more realistic and clarifying picture of the findings. Of all 15 categories, the most common structure overall is the noun phrase with *of*-phrase, representing in all cases more than 30% of all categories, with the highest frequency in the medicine RAs (35%). In particular, we found a total of 78 bundles with this structure, with a raw frequency of 375 –that is, LBs belonging to this category account for 32% of the total frequency of LBs in the corpus as whole. Biber et al. (1999) indicate that as much as 70% of the most common bundles usually consist of a noun phrase with an *of*-phrase fragment. The prevalence of this structure has also been found in previous studies on LBs (Chen & Baker, 2010; Hyland, 2008; Liu, 2012). As it could be expected given its high raw frequency, *the use of* is the most frequent bundle in this category (62 hits), with a higher presence in medicine RAs (21 hits). Other common examples are *one of the* (13 hits), *the analysis of (the)* (11 hits), and *the risk of* (11 hits). Examples (9), (10), (11) and (12) illustrate some of the most common LBs in this category:

(9) As the analysis of the selected linguistic features has illustrated, both adverbials and empty adjectives have been slightly more frequent in men's weblogs. (LIN_BD04_C)

(10) Disciplinary vocabulary also remains one of the most challenging areas. (LIN_RA12_I)

(11) That does not exclude the possibility of bias to the point where it is non-existent but it is an attempt to attenuate its effect. (MED_BD08_C)

(12) The use of different BMI reference values produced different prevalence estimates for the overweight category in the different populations. (MED_RA10_C)

The second most common structure is the other prepositional phrase, that is, bundles introduced by a preposition, excluding those with an embedded *of*-phrase; common LBs in this category are *of this paper, according to, in this study,* and *of the most*. We noticed above that LBs tend to be incomplete structural units; when they can be used as potentially complete units, these tend to act as discourse signaling devices (Biber et al., 1999, p. 999). The category of other prepositional phrases is one of the two which may integrate these complete structural units: see the examples from our corpus *between the two, as a result, on the other hand, in this (present) study/paper, on their own*; the other is the category of other noun phrases, e.g. *the present/current study, the following three.*

We have already mentioned particular examples of bundles which are especially recurrent in our corpus. One instance is *in order to*, which we consider a *to*-clause fragment (rather than a prepositional-phrase pattern; cf. Pérez-Llantada, 2014, for instance), and partly explains the relatively high frequency of the (verb/adjective +) *to*-clause structural pattern in all subcorpora. In addition, our data show two further common structures of bundles in specific subcorpora. One of them is the passive verb (+ prepositional phrase) with a higher use in the medicine RAs, exemplified by bundles such as *is associated with*, *have been proposed*, and *can be used to*, which interestingly are all found in the conclusion section of these texts. The impersonal nature of the passive construction seems to fit well with the medicine discipline, in which writers allegedly attempt to hide authorial interpretation more than their linguistics counterparts. This finding supports disciplinary differences on structural categories reported on in Hyland (2008, p. 11). The other structural category that shows a higher frequency than in other corresponding subcorpora is the noun phrase + verb phrase in BDs in linguistics. Examples of these bundles are *paper aims to*, *this paper will focus on* and *this study has.* We may hypothesize that this higher use is due to the emphasis placed on these non-agent text subjects in the teaching of academic discourse to university students.

A general tendency emerging from the figures represented in Table 4 relates to the variation in the use of LB structures. In this respect, RAs in linguistics show the greatest proportion of variation, as the only subcorpus illustrating all 15 categories. This subcorpus presents a rich range of different structural types of bundles, some of them of a more elaborated nature than in the learner corpus: e.g. the NP-based bundles *a growing interest in, our understanding of, body of research, avenues for future research,* and the VP-based

bundles *has the potential to, play an important role in.* Compared with this wide range of bundles, BDs in linguistics exhibit a less illustrative choice, with seven structural categories not represented, which can be explained by the less proficient writing skills of these authors. In the medicine corpora overall, however, the choice of bundles is definitely less varied. Curiously enough, medicine RAs show a much lesser degree of variation and representativeness in the use of LB structures, even though they belong to the same genre as their linguistics counterparts. It is difficult to say why this might be, but disciplinary variation and the topic of linguistic articles itself (language) could account for the discrepancies found.

The analysis of LBs according to discourse function has also revealed interesting insights. Table 5 provides an overview of the LB functions across genres and disciplines. As can be seen, bundles with text-oriented functions are prevalent over the other two types in general. The second most common type of bundle are those performing research-oriented functions. The comparison between these two functional categories, however, provides an interesting disciplinary distinction: whereas in linguistics there is a significant difference in frequency between the text-oriented and research-oriented functions in both learners and experts, and a particularly high use of text-oriented bundles (over 50%) in BDs, in medicine, on the other hand, the figures are closer between these two functions, and in medicine BDs they are exactly the same. This is (partly) in line with Hyland (2008, p. 14), who found a greater use of bundles with a referential function in the hard sciences to the same use in the soft-knowledge fields (i.e. linguistics), providing to the former "a greater real-world, laboratory-focused sense to writing", and thus emphasizing the empirical over the interpretative, as seen above. The more evident prevalence of text-oriented bundles in linguistics would also agree with this picture.

Table 5. Frequency of LBs per function: overall figures per genre and discipline (%)

Subcorpus	Research-oriented	Text-oriented	Participant-oriented	Totals
Linguistics BD	39.3	54.5	6.0	100
Linguistics RA	37.4	46.7	15.9	100
Medicine BD	42.4	42.4	15.1	100
Medicine RA	40.0	42.5	17.5	100

LBs with a participant-oriented function are far less frequent in our data with frequencies around 15%, except for the linguistics BDs, where the figure drops to only 6%. This underuse of participant-oriented bundles in our Spanish L1 writers agrees with findings

in other studies that have noted an avoidance of stance bundles in learners in comparison with English L1 authors (see Hyland, 2008, p. 19; Pérez-Llantada 2014, p. 91; Sheldon, 2018, p. 34). Pérez-Llantada (2014) notes that Spanish-speaking learner writers in English avoid personal markers to a greater extent than the corresponding expert writers of academic discourse. Our results also point to a lack of confidence on the part of the linguistics learners to express their stance and subjectivity.

In order to turn now to a more detailed analysis, we present Table 6 below with the figures of bundle types for the specific discourse functions included in each of the broad functional categories just mentioned. As with the discussion of the structure of bundles, a first thing to note is the greater and richer variety of functions in the linguistics RAs, with all ten categories represented in the table, in comparison with the other three subcorpora.

Table 6. LBs functions and their subcategories (%)

LBs functions		BDs		RAs	
		LIN	MED	LIN	MED
RES	Location	6.0	-	4.7	-
	Procedure	15.1	3.0	13.6	7.5
	Description	6.0	18.1	6.8	25.0
	Quantification	12.1	21.2	12.2	7.5
TEX	Transitions	-	3.0	4.0	2.5
	Resultative (inferential)	9.1	21.2	8.1	10.0
	Structure (identify/focus)	27.2	-	21.7	20.0
	Framing	18.1	18.1	12.9	10.0
PAR	Stance (probability, evidentiality, attitude)	6.0	15.1	14.9	17.5
	Engagement	-	-	1.0	-
Total		100	100	100	100

Concentrating on the most important functional category, that of text-oriented bundles, we see a clear preference for the structuring type in linguistics, and especially in linguistic BDs. Although the expert writers in medicine also exhibit an important use of this category, their learner counterparts, by contrast, make no use at all of these bundles, clearly preferring bundles with a resultative/inferential function instead, as will be discussed below. Structuring bundles, having an identifying and focusing meaning, allow writers to draw the reader's attention to a particular idea in the text, and to intensify the force of their arguments. Linguistics experts have used structuring bundles in their conclusions more often, a practice

which contrasts with their learner counterparts. These functional categories of bundles tend to be expressed by NP-based (common examples include *the aim of*, *the importance of* and *the current study*), as in (13) and (14), or VP-based structures (*aim of this paper is*, *this paper will focus on*, *there is a* and *that they are*), as in (15). The word *aim*, as noun or verb, is a recurrent one in bundles with this function.

(13) The aim of the present paper is to study the preference for the use of one-word verbs to multi-word verbs (LIN_BD09_I)

(14) This observation is consistent with the importance of cell-cell and cell-matrix contact in the activation of fibroblasts. (MED_RA25_C)

(15) This qualitative study has offered a general overview of those discourse functions which academic speech and writing have in common and those for which there is a marked difference in distribution. (LIN_RA05_C)

As just mentioned, resultative bundles are fairly common (21.2%) in medical BDs, by comparison with the other three subcorpora (with less than half this frequency), and by contrast, no instance of the structuring function was found. Interestingly, these writers have placed almost all their resultative bundles in the conclusion sections, as illustrated in (16) and (17). Other common bundles with this function are *the conclusion that*, *as a result of,* and *due to the fact that.*

(16) (…) call for the involvement of mental health professionals in the Emergency Room in order to offer a more complete evaluation of patients once medically stabilized. (MED_BD08_C)

(17) The results of this study demonstrate a need to distinguish at least two separate age-groups (…) (MED_BD10_C)

A final point worth mentioning in relation to the text-oriented category is that framing is more frequent in the learner corpus than in the expert data, exemplified by bundles such as *according to*, *related to the* and *as well as*. The greater need for these learners to situate

arguments with respect to others may have a genre-specific explanation; academic writing instruction may emphasize this writing strategy over others.

In research-oriented bundles, the second most important functional category, an interesting tendency arises: whereas the medicine data overall favor bundles contributing to the description of research objects, especially in RAs, linguistics favors the procedural bundles. This is not entirely surprising, considering the nature and object of study of each of these academic texts. And thus, whereas in medicine the description of the 'real-world' problem (medical conditions, clinical studies, etc.) is of great importance to their studies, in linguistics texts it is important to show the procedures of the research methods and demonstrate a certain ability in explaining how the research has been conducted. Both functions, i.e. method and procedure, are overwhelmingly often expressed by a NP-based bundle and very frequently by the noun phrase with *of*-phrase. Common bundles of description from the medicine texts are *the prevalence of, the presence of, the risk of,* and from the VP-based pattern, *it is a/the*. To express procedure, the most commonly used bundle is, by far, *the use of*. Other common bundles expressing procedure are *(the) analysis of (the)*, *the role of*, *the ways that,* and from the VP-based group of bundles, *can be used to*. Description and procedure bundles are exemplified in (18) and (19) respectively:

(18) Musculoskeletal disorders represent a relevant part of global morbidity and have an important impact on the prevalence of chronic diseases. (MED_RA03_I)

(19) This paper has tried to provide an accurate analysis of the English language in terms of lexical and grammatical parameters (…) (LIN_BD07_C)

A final insight from the group of research-oriented bundles is the high proportion of bundles with the meaning of quantification in medicine BDs, with respect to the other three corpora, and which again are mainly from the NP-based group of bundles. Examples include *the rest of, of the most* and *the most prevalent*.

The final category, participant-oriented bundles, mostly covers stance markers expressing opinion rather than facts, and may indicate degree of probability and epistemic meaning, on the one hand, or be part of the so-called 'other stance markers' (see Cortes, 2004, p. 209), on the other, which include LBs with evidential meaning, indicating the source of the information (e.g. *recent studies have*, *have been proposed*). The former type, the most common one, tends to be expressed by a recurrent set of structural categories, namely

anticipatory *it*-constructions containing an evaluative element (*it is true that, it would be interesting to, it can be argued that*), bundles with modal or semi-modal verbs (*should not be, seems to be*), epistemic adverbs, notably *likely* (*are likely to, is likely to be*), and other bundles expressing stance (*still in its infancy*, *has the potential to*). It is worth mentioning that stance can also be expressed in other ways than 3-, 4- and 5-word bundles, and that our study refers only to stance expressed in these sequences. Interestingly, stance is more common in the conclusion sections of the BD genre, whereas RAs contain more bundles of this type in their introduction sections: persuading readers from the very beginning through evidential and epistemic bundles seems to characterize more confident writing. Finally, engagement is almost non-existent in our corpus with only one bundle, namely *our understanding of,* used in the conclusion section of RAs in linguistics.

Conclusion

This paper has analyzed the use of LBs in the introduction and conclusion sections of learner and expert academic writing in linguistics and medicine. The quantitative and qualitative analysis performed in order to explore the frequency, structures and functions of LBs has yielded interesting results: LBs are very useful devices for the construction of discourse, but they behave in dissimilar ways in different disciplines and genres.

Regarding frequency, of the 218 bundles retrieved, 3-word bundles were more frequent in all subcorpora; of these, *the use of, in order to,* and *as well as* stand out as the most popular LBs. BDs in linguistics and medicine have produced a similar quantity of LBs in both sections, whereas RAs vastly differ in their frequency of use of LBs, which points towards a disciplinary difference. When comparing the learner and the expert corpus, on average, BDs in medicine contained more LBs than RAs in the same discipline, and the opposite tendency was found for linguistic BDs, which contained fewer LBs than their expert counterparts. In addition, a list of 35 convergent bundles was found, which can be a pedagogically useful resource for general academic writing. This quantitative analysis was complemented by qualitative analyses of structure and function which, after manual classification and revision of concordance lines, provided a more comprehensive picture of LB usage.

In terms of structure, both learner and expert writers favored NP-based bundles; the structure noun-phrase with *of*-phrase was by far the most frequent one in all corpora. BDs and RAs also agreed on the second most common LB structure: other prepositional phrase, which

allowed writers to include frequent discourse signaling devices in their texts. The main difference, however, lies in the greater structural variation of the LBs used by experts in linguistics; LBs in medical RAs, and in the learner genre, were definitely less varied. Finally, with regards to function, LBs performing text-orienting functions were the most prevalent in all subcorpora. The second group, LBs with research-oriented functions, was more popular among medicine expert writers, who seem to emphasize the empirical over the interpretative. The last function, participant-oriented, was the least represented one; this low frequency is especially marked in BDs in linguistics, which points towards a case of underuse. Additionally, while learners placed stance markers mostly in the last section of their texts, expert writers showed a preference for the use of stance in their introduction sections. Placement of LBs in particular sections of a text is yet another important feature that depicts writers' academic literacy. On the other hand, the lack of structuring bundles in medical BDs, and their recurrent use of resultative bundles also calls for explicit pedagogical attention. Disciplinary differences were also found regarding the prevalence of descriptive bundles in medicine, and of procedural bundles in linguistics; disciplinary conventions and the object of study of each of these texts could account for the discrepancies found.

The present study has some limitations worthy of mention. The first one is a methodological limitation: in order to extract sequences of words automatically, our retrieval method only included LBs that were fixed in nature; that is, our lists do not include variable bundles or bundles with open slots (e.g. *in section (...)*, *up to (...) %*, *to a (...) extent*). This method therefore does not capture LBs in their entirety. Including this type of permutations (e.g. using the ConcGram function in Wordsmith tools) could have helped to show a more comprehensive picture of LBs in academic writing (see O'Donnell et al., 2012). Another methodological limitation has to do with the fact that the learner corpus had not been error-tagged, which could have somehow affected the number of LBs extracted (i.e. if there were typos in particular words that were part of LBs, the software did not retrieve them). All texts included in the learner corpus, however, were successful BDs evaluated by their supervisors and the evaluating committee, so the probability of containing numerous typos is unlikely. Using a larger learner corpus would also have made the findings more representative. In addition, our analysis has looked at the use of LBs in the introduction and conclusion section of academic texts, as these sections tend to be the most conventional ones in these particular genres. Analyzing LB positions, not only with regards to sections but also with regards to paragraphs or sentences, would be interesting (see Römer 2010). Finally, when comparing our findings across previous studies that utilized corpora of different lengths and breadths, it was

difficult to accurately match the results. This limitation has also been attested to by Chen and Baker (2010, p. 43), who claim that "it is virtually impossible to find different corpora, of exactly the same size composed of the same number of texts, for direct comparison"; therefore, the cross-study comparisons included in this paper have to be regarded with caution.

Our analysis has provided a comprehensive list of 218 different bundles that may assist L2 learners to accommodate their academic writing to their specific discipline and genre. The results underline the importance these expressions have in order to write successful academic texts and to achieve disciplinary competence. As it has been shown, even though LBs are very frequent in language, mere exposure is often not enough for the acquisition and mastery of these devices in academic writing. Our findings therefore emphasize the need for more explicit teaching of LBs, always through corpus-informed materials, in agreement with the discipline and the genre studied.

Notes

[1]In the examples, the following abbreviations are used: MED (short for medicine) or LIN (linguistics) indicates the discipline, BD (short for Bachelor Dissertation) or RA (Research Article) indicates the genre, and I (short for introduction), or C (conclusion) indicates the section in which the LB was found. The number is the identification number assigned to each text.

[2]Linguistics journals selected were the following: *Applied linguistics; Computer Learner Corpora; Second Language Acquisition, and Foreign Language Teaching; Corpora and Language Teaching; English for Specific Purposes; Journal of Second Language Writing; Language Teaching Research; Lingua; Linguistics and the human sciences; TESOL Quarterly; Text: Interdisciplinary Journal for the Study of Discourse.*

Medicine journals were *BMJ Quality & Safety; European Journal of Clinical Investigation; Journal of international medical research; Journal of investigative medicine; Journal of the Canadian Association of Emergency Physicians; Lancet Neurol; Nursing Older People; Regenerative Medicine; The new England Journal of Medicine; Tissue Engineering.*

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. DOI: https://doi.org/10.1016/j.jeap.2013.08.002

Ädel, A., & Erman, A. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: a lexical bundles approach. *English for Specific Purposes 31*(2), 81-92. DOI: https://doi.org/10.1016/j.esp.2011.08.004

Anthony, L. (2018). AntConc (Version 3.5.7) [Macintosh OS X]. Tokyo, Japan: Waseda University. http://www.laurenceanthony.net/

Appel, R., & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: Differences between high-and low-proficiency levels. *Language Assessment Quarterly 13*(1), 55-71. DOI: https://doi.org/10.1080/15434303.2015.1126718

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes 26*(3), 263–286. DOI: https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. *Language and Computers 26*, 181-190.

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

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

Bondi, M. (2010). Metadiscursive practices in introductions. Phraseology and semantic sequences across genres. *NJES 9*, 99-123.

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL 5*(5), 31-64.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology 14*(2), 30–49.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: examples from history and biology. *English for Specific Purposes 23*, 397-423. DOI: https://doi.org/10.1016/j.esp.2003.12.001

Durrant, P., & Mathews-Aydınlı, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes 30*(1), 58-72. DOI: https://doi.org/10.1016/j.esp.2010.05.002

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective* (pp. 27–49). Amsterdam: John Benjamins Publishing. DOI: https://doi.org/10.1075/z.139.07gra

Grant, L., & Bauer, L. (2004). Criteria for re-defining idioms: Are we barking up the wrong tree? *Applied Linguistics 25*(1), 38-61. DOI: https://doi.org/10.1093/applin/25.1.38

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes 27*(1), 4-21. DOI: https://doi.org/10.1016/j.esp.2007.06.001

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes 31*(1), 25-35. DOI: https://doi.org/10.1016/j.esp.2011.07.002

Meunier, F., & Granger, S. (Eds.). (2008). *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins Publishing. DOI: https://doi.org/10.1075/z.138

Mur-Dueñas, P. (2011). An intercultural analysis of metadiscourse features in research articles written in English and in Spanish. *Journal of pragmatics 43*(12), 3068-3079. DOI: https://doi.org/10.1016/j.pragma.2011.05.002

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nesselhauf, N. (2005). *Collocations in a learner corpus* (Vol. 14). Amsterdam: John Benjamins Publishing.

O'Donnell, M. B., Scott, M., Mahlberg, M., & Hoey, M. (2012). Exploring text-initial words, clusters and concgrams in a newspaper corpus. *Corpus Linguistics and Linguistic Theory 8*(1), 73-101. DOI: https://doi.org/10.1515/cllt-2012-0004

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. DOI: https://doi.org/10.1016/j.jeap.2014.01.002

Römer, U. (2010). Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction* 3(1), 95-119. DOI: https://doi.org/10.1075/etc.3.1.06rom

Sheldon, E. (2018). Dialogic spaces of knowledge construction in research article Conclusion sections written by English L1, English L2 and Spanish L1 writers. *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE) 35*, 13-40.

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

Sinclair, J. (2005). Corpus and text-basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1-16). Oxford: Oxbow Books.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Appendix 1

LBs found in the learner and the expert corpus according to sections and disciplines (sorted by frequency)

LIN BD introduction		**MED BD introduction**		**LIN RA introduction**		**MED RA introduction**	
in order to	14	in order to	7	the use of	17	the use of	10
the aim of	8	as well as	7	in order to	14	the risk of	7
of this paper (is to)	8	such as the	5	(used) to refer to	10	as well as	6
the analysis of (the)	7	according to the	5	in terms of	9	a number of	5
the use of	6	the rest of	5	refer to the	8	of this study	4
as well as the	5	the prevalence of	5	of the most	8	in order to	4
the fact that	4	the result of	4	the effects of	7	the effect/s of	4
(one) of the most	4	of the most	4	one of the	7	the presence of	4
in this paper	4	the use of	4	the basis of	6	been shown to	4
the study of	4	the risk of	4	as well as	6	to be the	4
it has been	4	the development of	4	some of the	6	it is not	3
(one) of the main	4	there is no	4	different types of	6	there is a	3
due to the (fact that)	4	is one of the	4	of the same	5	the prevalence of	3
to the study (of the)	4	the conclusion that	3	that they are	5	changes in the	3
paper aims to	3	as a result	3	the current study	5	the ability to	3
attention to the	3	of the population	3	the present study	5	be able to	3
related to the	3	lack of a	3	there is a	5	as a result (of)	3
to do so	3	the most prevalent	3	based on the	5	**MED RA conclusion**	
followed by the	3	is the most	3	in the field	5	the use of	11
there is a	3	it is a	3	the nature of	5	the current study	9
aim of this paper is	3	it is the	3	are likely to	5	as well as	7
this paper aims (to)	3	is not a	3	the comparison of	4	is associated with	7
this paper will (focus on)	3	recent studies have	3	between the two	4	in this study	6
LIN BD conclusion		have been proposed	3	interest in the	4	was associated with a/an	6
the use of	11	although there is (no)	3	in this study	4	a number of	6

in order to	8	**MED BD conclusion**		in this paper	4	the proportion of	6
the fact that	5	of this study	6	the focus of	4	the presence of	6
of this paper	4	in order to	5	the results of	4	the present study	5
as well as	4	the possibility of	5	the area of	4	the results of	5
most of the	4	due to the (fact that)	5	the context of	4	consistent with the	5
it has been	4	the results of	3	the fact that	4	can be used (to)	5
this study has	3	impact of the	3	the range of	4	in addition to	4
one of the	3	one of the	3	the role of	4	there was no	4
analysis of the	3	will not be	3	the ways that	4	the prevalence of	4
				can be used	4	in our study	3
				the field of	4	because of the	3
				in the current	4	the application of	3
				the notion of	4	the field of	3
				the study of	4	we did not	3
				that it is	4	are needed to	3
				it has been	4	there is a	3
				argue that the	4		
				in the context of	4		
				a (wide) range of	4		
				to contribute to (the)	4		
				differences in the	3		
				to find out	3		
				the importance of	3		
				in the study	3		
				focusing on the	3		
				as a result	3		
				in relation to	3		
				the number of	3		
				a number of	3		
				as part of	3		
				in a number of	3		
				to develop a	3		
				analysis of the	3		
				in what ways	3		
				is used to	3		
				understanding of the	3		
				the form of	3		
				body of research	3		
				the potential to	3		
				contribute to the	3		
				be argued that	3		
				is the use of	3		
				in the form of	3		
				a growing interest in the	3		
				on the basis of the	3		
				it can be argued that	3		
				of the use (of)	3		
				is likely to (be)	3		
				has the potential (to)	3		
LIN RA conclusion							
the use of	26	but it is	3				
the present study	10	that there are	3				
in this study	10	it is important	3				
in order to	10	study has shown	3				
the fact that	10	study has been	3				
in this paper	8	this study is	3				
there is a	8	is that the	3				
as well as	8	the part of	3				
seems to be	8	reference to the	3				
the case of	7	in this way	3				

in the use of	7	the following three	3
in relation to	6	a variety of	3
in terms of	6	some of the	3
the lack of	6	the majority of	3
differences in the	6	the number of	3
of the most	6	be used to	3
in the present study	6	can be used to	3
the importance of	5	the construction of	3
the current study	5	the level of	3
based on the	5	the process of	3
with respect to	5	the role of	3
should not be	5	the beginning of	3
in the case of	5	for the present	3
this study has	4	found in the	3
this paper has	4	the complexity of	3
such as the	4	understanding of the	3
for future research	4	on their own	3
due to the	4	to be the	3
has shown that	4	it should be	3
greater use of	4	there is no	3
the analysis of	4	on the other hand	3
the quality of	4	avenues for future research	3
in the literature	4	on the part of	3
that there is	4	it is true that	3
needs to be	4	still in its infancy	3
our understanding of	4	play an important role in	3
it would be (interesting to)	4		

Authors' information
Noelia Navarro Gil is currently pursuing her PhD in English Linguistics at Complutense University of Madrid. She graduated in English studies at the same university, and holds an MA in TEFL from Universitat Pompeu Fabra. She is a member of the Institute for Multilingualism of Universitat Internacional de Catalunya, where she works as a professor of research methods for health sciences. Her main research areas include corpus linguistics, disciplinary literacy, second language writing, and more specifically, L2 learners' academic discourse.
E-mail: nnavarrog@uic.es

Elena Martínez Caro holds an MA degree in Linguistics from the University of Reading (UK) and a PhD from Complutense University of Madrid, where she is currently associate professor in English linguistics. Her main areas of research are the syntax-pragmatics interface (information structure), discourse analysis (discourse segmentation and discourse markers) and English grammar, including its contrasts with Spanish. In these areas she has published a book, several articles and book chapters and, recently, with Lachlan Mackenzie, a grammar of English for Spanish speakers (*Compare and Contrast*. Comares, 2012).
E-mail: elenamc@ucm.es

To cite this article:
Navarro Gil, N. & Martínez Caro, E. (2019). Lexical bundles in learner and expert academic writing. *Bellaterra Journal of Teaching & Learning Language & Literature*, *12*(1), 65-90. DOI: https://doi.org/0.5565/rev/jtl3.794