Article

LANGUAGE
TEACHING
RESEARCH

Language Teaching Research
1–25
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1362168819877306
journals.sagepub.com/home/ltr
SAGE

Academic lexical bundles in graduate-level math texts: A corpus-based expert-approved list

Abdullah Alasmary
King Saud University, Saudi Arabia

Abstract

The purpose of this article is to synthesize and analyse the most frequently occurring, widely dispersed and pedagogically useful lexical bundles in mathematical texts. Drawing on a five-million-word corpus of graduate-level textbooks, a total of 65 academic sequences meeting a predefined set of length, frequency and distribution criteria are obtained for functional and structural analyses. Results indicate that the structural forms of the sequences are clausal and that the dominant function is research-oriented. Sequences are rank-ordered according to a composite score obtained as a result of triangulating frequency profiles, distribution proportions and expert judgements. This article concludes by highlighting the pedagogical implications of the study for both language educators and mathematics practitioners.

Keywords

lexical bundles, mathematical discourse, pedagogic materials, recurrent linguistic patterns

I Introduction

The prevalent presence of multiword sequences in academic speech and writing has drawn much research during the past few years (Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004; Hyland, 2008a). While known by different terms, including lexical bundles (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Chen & Baker, 2010; Pan, Reppen, & Biber, 2016), multi-word constructions (Liu, 2012), academic formulas (Simpson-Vlach & Ellis, 2010), formulaic language (Pérez-Llantada, 2014), and word combinations (Ädel & Erman, 2012), these sequences are key for language production and comprehension and their mastery is perceived as a sign of academic maturity and

Corresponding author:
Abdullah Alasmary, Department of English Language & Translation, College of Languages & Translation, King Saud University, Riyadh, 11451, Saudi Arabia
Email: aasmary@ksu.edu.sa

group membership. With the proliferation of corpus tools and automated software programs, researchers are now able to examine texts for patterns of lexical bundle use, revealing insights into how often these patterns recur and the extent to which they spread across texts.

Previous research into academic formulas has focused on the structural, functional and distributional characteristics featuring recurrent sequences across a wide range of academic disciplines (e.g. Biber et al., 2004; Hyland, 2008a; Simpson-Vlach & Ellis, 2010). As evidence emerges that these patterns tend to be register-specific (Hyland, 2008b), researchers begin to take a different approach, thus focusing on single domains or specific discourse settings such as biology and history (Cortes, 2004), pharmacology (Grabowski, 2015), law (Breeze, 2013), telecommunications (Pan et al., 2016), EU documents (Jablonkai, 2010), and applied linguistics (Qin, 2014). These studies demonstrate that while there are few academic expressions which transcend discipline boundaries, several expressions are register-specific.

While there have been several attempts to investigate aspects of recurrent patterns and repeated vocabulary use in a range of disciplines, mathematics appears to have received little scrutiny. Liu (2012, p. 25) highlights the need to examine underexplored areas, stating that it is ‘helpful for language learning/teaching to identify those most frequently used multiword constructions in various registers’. To the best knowledge of the researcher, there has been no previous attempt to produce a list of such multiword patterns in mathematics; thus, this article aims to produce a mathematics-oriented list of sequences combining the power of corpus tools and the opinions gleaned from specialists. Although the list is compiled to be easily accessible to a wider readership, students who would like to pursue graduate education in mathematics and for whom English is a second/foreign language will find it useful for fostering their knowledge of discipline-specific vocabulary. Also, materials designers, textbook authors and teaching professionals may also draw on the components of this list for insights on which vocabulary to prioritize, given the scarcity of time and resources that shapes several instructional settings.

I Mathematics, language and lexical bundle research

The role of language in the construction, dissemination and interpretation of mathematical knowledge is undeniably great, as is pointed out by both mathematics educators and applied linguists. Mathematics educators have explored a variety of topics at the intersection between language and mathematics, including the use of formal versus informal vocabulary instruction (Leung, 2005), mathematical concepts and the way they are defined in different discourse settings (Morgan, 2005), the numerous linguistic problems facing non-English-speaking mathematicians (Chan, 2015; Schleppegrell, 2007), the situated characteristics of mathematics as a discipline (Brilliant-Mills, 1994) and the multi-semiotic nature of mathematical register (O’Halloran, 1998, 2005, 2015). Within applied linguistics and English for specific/academic purposes, interest in the study of mathematics has focused on a limited range of topics, including the investigation of the stance and engagement patterns in pure mathematics research articles (McGrath & Kuteeva, 2012), the rhetorical structure of the introduction sections in journal articles published in the domain of mathematics (Graves, Moghaddasi, & Hashim, 2014), the organization and content of research articles in pure mathematics (Kuteeva & McGrath, 2015), and the processes underlying an online collaboration of authors to draft, revise and publish research articles in pure mathematics (McGrath, 2016).

Two key studies have particularly focused on lexical bundles in mathematics. The study by Herbel-Eisenmann, Wagner and Cortes (2010) was conducted to identify stance bundles occurring in a corpus of classroom interactions. The analysis of stance patterns showed that these patterns served to signal desire (e.g. *if you want to*), obligation (e.g. *you don't have to*) or intention (e.g. *you are going to*). Another exploration of mathematical discourse was carried out by Cunningham (2017) who extracted frames with fillable slots from a corpus of journal articles and subjected them to structural and functional classification. The structural forms of the frames were then grouped into noun, preposition and verb categories, following similar structural patterns identified in several previous studies. Functionally, the frames were categorized into two key groups: aboutness (e.g. *the proof of* *) and coherence (e.g. *in this section* *).

2 Bundle listings

The interest in the study of lexical bundles has encouraged researchers to create several bundle lists, most of which are built with a recommendation for pedagogical use (Ackermann & Chen, 2013; Durrant, 2009; Hsu, 2014; Liu, 2012; Martinez & Schmitt, 2012; Shin & Nation, 2008; Simpson-Vlach & Ellis, 2010; Wood & Appel, 2014). One of the earliest attempts to suggest a list of recurrent clusters was made by Shin and Nation (2008) who focused on patterns occurring in the spoken section of the British National Corpus (BNC). The scope of the study was wide ranging, covering bundles of different lengths and types: two words (e.g. *you know*) three words (e.g. *at the moment*) four words (e.g. *thank you so much*) and frames with fillable slots (e.g. *it seems (N, A, to INF, that SV)*). The study by Wood and Appel (2014) was intended to explore the most recurrent lexical bundles in 1.6 million-word corpus of first-year engineering and business textbooks. The final elements on the list were then searched in a 15,839-word corpus of textbooks for English for academic purposes (EAP) to determine whether common EAP textbooks paid attention to these bundles. A major finding was that a great number of bundles obtained from the disciplinary corpus did not appear in the comparison corpus, an outcome that indicated poor representation of these key clusters in the academic texts. A further examination of the bundles occurring in the specialized corpus revealed that none of them 'is being presented as teachable units, or highlighted for the reader in any significant way' (Wood and Appel, 2014, p. 8). On a larger scale, Hsu (2014) applied a set of criteria such as semantic opaqueness and non-compositionality to derive recurrent bundles from a 20-million word corpus of college textbooks, representing 40 different academic domains. A list of 475 sequences was retrieved and then subjected to structural and functional analysis. Bundles took different structural forms, including phrasal verbs (e.g. *account for, cope with*), passive fragments (e.g. *be accustomed to*), and prepositional phrases (e.g. *in the light of*). Functionally, the academic constructions were found to serve as referentials (72%), discourse organizers (22%), and stance expressions (6%).

Another key list was compiled by Ackermann and Chen (2013) who used a combination of corpus methods and expert intervention to build a list of collocations representing a range of academic disciplines. The final list included several patterns: verb plus noun (e.g. *cast doubt*), adjective plus noun (e.g. *academic writing*), noun plus noun (e.g. *background knowledge*), verb plus adjective (e.g. *make explicit*) and adverb plus adjective (e.g. *highly controversial*). A closer look at the distribution of collocations revealed that the greatest number of these patterns were made of nominal collocations, whereas adverb+adjective patterns formed the smallest category. Liu (2012) compiled a similar list incorporating multiword constructions, another name for lexical bundles, from previous research and looked at their distribution in two large corpora: the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC). The master list includes a total of 228 bundles, most of which are phrasal and tend to be domain-specific. Functionally, patterns fall into one of three categories: referentials (e.g. *in terms of the*), stance markers (e.g. *we argue that the*) and discourse organizers (e.g. *as a result of*). The Phrasal Expressions List, compiled by Martinez and Schmitt (2010), is another attempt to generate a list of recurrent patterns for pedagogical use. By applying several criteria, such as frequency of occurrence, meaningfulness and non-compositionality, authors aggregated a total of 505 phraseological patterns derived from the British National Corpus. A final, and by far the most methodologically robust, attempt to suggest a list of academic bundles was conducted by Simpson-Vlach and Ellis (2010), who applied a combination of corpus-based and human intervention criteria for identifying potential bundles in academic speech and writing. Instead of reporting each of these characteristics separately, the authors adopted an innovative way in which bundles were arranged according to a composite score-Formulaic Worthy of Teaching (FWT). While the list is compelling, thus incorporating patterns of great utility for learners, it leans heavily toward general academic use, not discipline-specific one.

While all these lists are of greater pedagogical benefit which language instructors can draw upon in teaching as well as in preparing teaching materials, it is important to note that they are not without problems. First, and with the exception of the list generated by Wood and Appel (2014), the focus of these lists is on several domains and registers, a procedure that carries the risk of sidelining some core, discipline-based recurrent patterns typical of specific academic domains. A second problem relates to the concern raised by Hyland & Tse (2007) who questioned the merits of creating a one-size-fits-all list that could be used irrespective of the target domain or discipline. Third, the items in these lists are not comprehensive enough to account for recurrent patterns in some underexplored disciplines such as mathematics. This study thus aimed to address this vacuum by exploring a corpus of textbook materials for patterns of repeated vocabulary use. To direct the process of this research, answers to the following questions are sought:

1. Which lexical bundles are the most recurrent and the most widely spread in advanced mathematics textbooks? What are their structural and functional attributes?
2. Which lexical bundles are of the greatest utility for EAP students, as attested by expert mathematicians?

II Methodology

In this section, I will discuss the content of the corpus and its construction, the processes involved in the extraction of the target bundles and the procedures undertaken to refine and distill the corpus-derived data. A discussion of the qualitative part of the study, including the steps followed to elicit the opinions of experts on the usefulness of bundles and the discourse functions performed by the selected bundles, will conclude this section.

1 The construct of pedagogical utility

Given that the key purpose underlying this study is to generate a list of pedagogically useful vocabulary for language-learning purposes, it is important to select items on the list based on sound methodological considerations, as is summarized below:

- Genre: The list is discipline-specific and register-focused. It draws on written academic textbooks aimed at graduate students in the domain of mathematics.
- Bundle length: This study addresses four-word bundles because they are, as pointed out by Hyland (2008b, p. 8), 'far more common than 5-word strings and offer a clearer range of structures and functions than 3-word bundles'.
- Frequency of occurrence: A first step in the compilation of the list involves applying frequency of occurrence, a key statistical measure that has long been used in listings of key academic vocabulary (e.g. Coxhead, 2000). Martinez and Schmitt (2008, p. 302) maintain that frequency of occurrence is 'one of the best indicators of usefulness of individual words in general English'.
- Range: Some researchers suspect that the frequent use of an item reflects a purely idiosyncratic use characteristic of a specific author or text (Cortes, 2004). Range is used alongside frequency to ensure that the selected item is widely used across several texts making up the entire corpus.
- Expert ratings: A final step in the process of compiling the list involves seeking the opinion of professionals regarding the usefulness of statistically derived items for mathematics students in general and those pursing graduate work in mathematics in particular. This procedure has been pursued by researchers (e.g. Simpson-Vlach & Ellis (2010) who argued that human intervention is needed to distill data.
- Ranking: All items on the list are presented based on a combination score representing frequency of occurrence, range and the opinions of experts.

2 Study corpus

The corpus used in this study consists of 36 full-length textbooks from a series of publications aimed for graduate students in mathematics (for the full list of texts, see Appendix 1). The texts cover a wide range of mathematical topics, including enumeration, algebra, probability, calculus, number theory and matrices, to mention but a few. In selecting these textbooks, no decision is taken to differentiate between pure and applied mathematics, as a distinction of this type 'is often difficult to make and lacks a clearly argued epistemological basis' (Graves et al., 2014, p. 2). Prior to corpus treatment, texts are cleared of acknowledgements, tables of contents, lists of references, and indexes. Unlike in Wood and Appel (2014), activities and exercises sections, whether embedded in the text or supplied at the end of chapters or sections, are not removed, given their status as part of the text. Each textbook is given a distinct code so as to facilitate concordance checks needed to refine data and determine the functions of lexical patterns. Textbooks are then converted into plain texts, allowing for their treatment using corpus-automated tools. The Cluster Function in the WordSmith program (Scott, 2016) is used to extract recurrent patterns meeting the predetermined length and frequency scores.

3 Sequence identification

The identification process necessitates determining the number of words included within a single bundle, the frequency of the target bundles and their distribution across the texts making up the corpus. As for the length of the target sequence, there appears to be a tendency in lexical bundles research to target patterns incorporating four words (Cortes, 2004). With respect to the frequency of occurrence, there is no consensus among scholars on a specific frequency cut-off score beyond which a specific bundle can be included for analysis. In the literature, frequency cut-off scores vary from one study to the other, ranging from 10 (Simpson-Vlach & Ellis, 2010) to 50 times per million words (Breeze, 2013). Given these two extreme frequency thresholds, a rather moderate frequency score of 25 times per million words was adopted. A third consideration in the selection process of bundles concerns the number of texts in which the target expression should occur before being included in the final list of target bundles. The impetus behind this procedure is to ascertain that the recurrent pattern is not the result of an idiosyncratic use typical of a particular author or text (Biber & Barbieri, 2007). Some researchers have adhered to a minimum number of texts (e.g. Biber et al., 2004), while others consider measuring distribution according to a specific proportion (e.g. Hyland, 2008b). Given that the goal underlying this study is to offer mathematicians and specialists access to a vocabulary list with greater utility, a lexical bundle is included for analysis if it is found in 75% of the texts making up the corpus. This conservative distribution score aims to ensure that bundles appearing in the list are in wide use by different authors and are not a characteristic of a specific text or writer.

By applying these criteria, I am aware of the fact that a lot of potentially pedagogically useful sequences will not be included in the final listing simply because they fall below the predetermined distribution or frequency thresholds. However, this list is built with a goal that the candidate sequences will be incorporated into language teaching programs or used for creating pedagogic materials and including a large number of them will overwhelm students and overburden instructors. The process of applying length, frequency and dispersion parameters has resulted in an initial list comprising a total of 98 bundles.

4 Data refinement

The list obtained after applying length, frequency and range criteria are not without problems. An extensive process of checking concordance lines has shown a great deal of overlapping between similar bundles. Including overlapped bundles in a list may not be appropriate, given that one purpose underlying this research is to provide practitioners and math educators with various types of academic bundles, not variants of the same bundle. An example of such overlapping is found in sequences such as *and only if the, and only if it*, and *and only if there*, which are extensions of the bundle *if and only if*. Following the practice in previous research (e.g. Chen & Baker, 2010), closely related bundle types are combined into a single bundle. Another step to filter the list involves merging bundles that are identical except for one word, normally occurring in the bundle-final position. The bundles *that there is a, then there is a*, and *in this case the* are similar to the sequences *that there is an, then there is an* and *in this case we* except for the item occurring at the end of the sequence. To account for the existence of a closely related bundle form, the different word is separated from the original item by a slash (e.g. *that there is a/an*). Another refinement procedure addresses a group of lexical bundles which begin with the directive cognitive verb *let* and differ in a single variable occurring in the same position (e.g. *let x be a, let g be a*). This pattern is seen as a productive frame (Cunningham, 2017) which allows for various variables to be inserted within its boundaries. There are five *let*-initiated patterns with different frequency profiles, so the decision was taken to include the most recurrent pattern with an asterisk above it, indicating that other variables of the same frame also exist. Given the nature of the corpus under study, the retrieval process may entail academic sequences that are made of more symbols than words. Following McGrath and Kuteeva (2012) bundles including mathematical symbols or characters denoting mathematical entities are not included in the data (e.g. *a* and *b* are). The list after refinement incorporates a total of 71 bundles which will be further filtered after obtaining the opinions of math experts.

5 Expert opinion

In order to select items worthy of classroom attention from the corpus-derived list, it was decided to seek the opinions of 20 professionals with experience in publishing scholarly papers and materials in mathematics. Ackermann and Chen (2013, p. 236) have warned that ‘only with human intervention can a data-driven collocation listing be of much pedagogical use while still taking advantage of statistical information to help identify and prioritize the corpus-derived items.’ Experts handling the ranking process are chosen based on four criteria: expertise in mathematics education (years of teaching mathematics), publication record (they must have published a textbook, textbook chapter or a scholarly article prior to this study), academic rank (at least an assistant professor), and language of publication (English as a medium of publication). The participants include nine full professors, eight associate professors and three assistant professors. Their average length of experience is 25.55 years (Minimum 6, Maximum 43, SD = 11.06), and the average number of academic papers they have published is 35.10 (Minimum 2, Maximum 130, SD = 36). Besides scholarly papers, two participants have reported publishing eight textbooks (four texts each) and a third participant, seven texts. All experts have mentioned that they use English for publication with one of them adding French. The refined list of bundles resulting from the corpus treatment was given to these experts who were requested to evaluate items based on a 5-point Likert scale, in response to whether they think the sequence merits classroom attention and thus should be included in a list for learning purposes. The survey is presented to them with a brief explanation of the study goals, filling-in procedures, and the full list of sequences. Their responses range from extremely agree to extremely disagree:

1 = extremely disagree that the item is pedagogically useful;
2 = disagree that the item is pedagogically useful;
3 = not sure;
4 = agree that the item is pedagogically useful;
5 = extremely agree that the item is pedagogically useful.

The intra-class correlation coefficient is 0.73, showing moderate agreement among raters. A bundle was removed from the list if the sum of the expert judgments on it was under 80. The impetus for adhering to such conservative score is to ensure that items in the final list are of great pedagogical utility. Expert ratings have reduced the number of sequences to 65, thus eliminating six bundles deemed unworthy of classroom attention.

Expert opinions were also sought when determining the functions of the target bundles. The functional analysis was conducted by four professionals, none of whom took part in the survey measuring the usefulness of bundles. Here is a brief description of each expert handling the functional classification:

1. associate professor majoring in applied linguistics and English for specific purposes (ESP) with a long experience teaching English to Engineering students;
2. assistant professor majoring in English and having an extensive experience in teaching ESP to science students;
3. assistant professor majoring in English as a second language (ESL) and currently teaching English to non-English-majors, most of whom are science students;
4. full professor majoring in mathematics who had published extensively in various mathematical outlets.

A Cronbach alpha was conducted to measure inter-rater agreement, and it was found at .56. This moderate result is admittedly unsurprising, given the highly specialized nature of the math register. The panel decided to meet and, as a group, work out the discrepancies with the help of concordance lines. A unanimous agreement was reached and the bundles were functionally classified based on the group decision of the panel members.

III Results

The final list of lexical bundles (see Appendix 2) includes a total of 65 items, arranged according to a composite score representing the frequency of occurrence, distribution across texts and the expert ranking. It is clear that the bi-conditional bundle *if and only if* is by far the most recurrent, the most useful and the most widely dispersed sequence in the list. Although the bundle *on the other hand* is the most widespread and the

Table 1. Nominal lexical bundles.

Structural category	Grammatical pattern	Lexical bundle
NP-based	NP + *of*	1. the set of all 2. the proof of theorem 3. (as) + the proof of the 4. the definition of the 5. a finite number of 6. the existence of the/a 7. the sum of the 8. a linear combination of 9. an example of a 10. (is) + a consequence of the 11. a special case of
	Other NPs	1. the right hand side 2. the fact that the 3. the left hand side

second-most recurrent bundle in the corpus, it is perceived as less useful by experts, thus occupying a relatively low position in the ranking. Some lexical bundles are considered extremely useful by the experts although they do not exhibit higher frequency and distribution characteristics (e.g. *then there exists a*). In a similar way, some patterns that are widely spread across the corpus are not as frequent and expert-appreciated as expected. An example of this group is found in the sequence *the fact that the* which appears in 97 % of the texts, but demonstrates rather low values on the frequency and expert-ranking scales. Given these discrepancies in frequency, dispersion, and usefulness parameters, prioritizing bundles based on a composite score incorporating these values is more pedagogically compelling and conducive to better use in classroom. In this section I will first discuss the structural forms of the retrieved academic bundles, giving examples for each category from the corpus. Next, all bundles will be classified according to the functions that they perform in the discourse.

1 Structures of bundles

The literature is replete with several frameworks developed to classify bundles according to distinct structural forms (e.g. Biber et al., 2004; Chen & Baker, 2010; Cortes, 2004; Hyland, 2008a; Pan et al., 2016). Following that tradition, all items in our list are assigned to three major groups: noun-, preposition- and verb-based bundles. A fourth group is created to account for fragments that do not fall neatly into any of the three structural categories. As can be seen in Table 1, the first group of bundles comprises noun phrases, the largest number of which include embedded *-of* constructions. Some nominal bundles are preceded by the definite article *the* such as *the proof of theorem* and *the sum of the*, while others take the indefinite article *a* or its variant *an* (e.g. *a linear combination of* and *an example of a*). For the sake of space, I will mention two examples illustrative of each structural type.

Table 2. Prepositional lexical bundles.

Structural category	Grammatical pattern	Lexical bundle
PP-based	PP + *of*	1. in the proof of 2. in the case of 3. in terms of the 4. from the definition of 5. in the definition of
	Other PPs	1. in this section we 2. with respect to the 3. in this chapter we 4. from the fact that 5. on the other hand 6. in the next section 7. in this case the/we 8. in such a way

1. **The proof of Theorem** 9.2 uses change of ring constructions.
2. We next consider **an example of a** finitely generated discrete subgroup of M(B3) that is not geometrically finite.

The second category includes preposition-headed sequences, almost half of which end with an *-of* pattern (Table 2). The remaining set of sequences begin with a preposition and conclude with either a marker for the beginning of another clause (e.g. *in this section we*) or a post-modifier (e.g. in the case the/we). The following are examples of sequences in this group:

3. **In the case of** sesquilinear forms, this can happen provided the field automorphism extends to the larger field, i.e. if k is odd.
4. **With respect to the** right half plane, the subset of the imaginary axis {is : |s| > λ} has harmonic measure h(w) = 1

In contrast to the two nominal categories, the third structural type consists of constructions comprising a verb component. These patterns can be further subcategorized into active/passive verbs, copula-verb-based constructions, existential *there*, and *it*-clauses. Apparently, the use of the verb *be* appears to dominate this group, as it makes presence in several subcategories. The first group of verb-based sequences consists of copula-be verb followed by noun, adjective, or prepositional phrases:

5. An easy application of Theorem 6.2 **is the set of** Weyl's inequalities.
6. The condition that Ap(¢) # 1 **is equivalent to the** condition that the fixed point P has multiplicity 1, so Theorem 1.14 holds provided ¢ has d + 1 distinct fixed points.

As can be seen in Table 3, other verbal subgroups incorporate existential '*there*', *it*-clauses, and active/passive constructions. Here are examples of each type:

Table 3. Verb-based lexical bundles.

Structural category	Grammatical pattern	Lexical bundle
VP-based	Copula *be* + NP/AP	1. be the set of
		2. is the set of
		3. is a set of
		4. is an element of
		5. is equivalent to the
		6. is the same as + (the)
		7. is equal to the
		8. is of the form
		9. is independent of the
	Existential phrase	1. that there is a/an
		2. there is a unique
		3. then there is a/an
		4. then there exists a
		5. that there exists a
		6. there exists a unique
	VP with active verb	1. we may assume that
		2. let (*x*)* be a
		3. to show that the
		4. show that there is
		5. we see that the
		6. we have the following
		7. we say that a
		8. (it) + suffices to show that
	It-clauses	1. it is easy to + (to see that)
		2. it is enough to
		3. it follows from the
		4. it follows that the
	Passive verb + PP fragment	1. is said to be
		2. can be written as
		3. is contained in the
		4. is given by the
		5. can be used to
		6. can be found in

7. The conclusion, then, is that for any $E < 0$, **there is a unique** (up to a constant) solution to (5.2) that is square-integrable on the interval $(-\infty,-A)$.
8. **It is easy to** find level sets of smooth functions that are not smooth sub-manifolds.
9. In particular, any complex function on G **can be written as** a linear combination of the character.

Table 4 includes the rest of the bundle structures which are mainly conditionals and fragments. Patterns with the conditional *if* are very frequent in the data, the most recurrent of which is the form *if and only if*:

10. In this section we work over an algebraically closed base field k, so that a scheme X is smooth **if and only if** it is nonsingular.

Table 4. Other structural types of lexical bundles.

Structural category	Grammatical pattern	Lexical bundle
Other structures	Conditionals	1. if and only if + (the)/(it is)/(there) 2. if there is a
	Fragments	1. such that for all 2. and the fact that 3. that the set of

2 Functions of the bundles

Another key purpose underlying the present study is to categorize bundles according to the functions that they serve in the texts. The functional framework developed by Hyland (2008a), which was subsequently applied by other researchers (e.g. Durrant, 2017; Pan et al., 2016), is chosen over that of Biber et. al (2004) because an initial piloting reveals that Hyland's framework is comprehensive enough to account for the numerous discoursal functions exhibited by the sequences in the list. This is not surprising, given that the framework was first used to put groups sequences commonly found in academic writing into functional groups. According to this functional scheme, recurrent clusters are divided into three main groups: research-oriented, text-oriented and participant-oriented. Bundles serving a research-oriented function 'help writers to structure their activities and experiences of the real world' (Hyland, 2008a, p. 49). This functional group consists of several bundles that signal location, procedure, quantification, description, and topic. The second major category involves text-oriented bundles that are 'concerned with the organization of the text and its meaning as a message or argument' (2008a, p. 13). Bundles in this category are classified into transition, resultative, structuring, and framing signals. The third functional group includes participant-oriented bundles which serve to highlight the range of attitudes, opinions or judgements expressed by the writer. Bundles in this group fall into two major categories: stance and engagement.

The most bundles in the list perform a research-oriented function (see Appendix 2). This is not uncommon, given that one of the most important communicative purposes of graduate textbooks is to disseminate content knowledge to a presumably less-informed audience. Bundles in the research-oriented group can be further divided into subgroups, thus serving to mark existence/uniqueness, allude to location, refer to a topic-related item or procedure, define mathematically-oriented concepts, and describe an attribute or elaborate on a procedure or operation. Existence/uniqueness markers include eight bundles: *that there is a/an, there is a unique, then there exists a, then there is a/an, that there exists a, there exists a unique, show that there is* and *the existence of the/a*. These patterns are used to allude to some mathematically specific entities, such as a property, subset, function or form. Here are two examples illustrating existence markers:

1. We then consider the Berezin transform of a function and show **that there is a** semigroup property with respect to the parameter α.
2. If f is a continuous open map of a locally compact Hausdorff space X onto a Hausdorff space Y and if K is a compact subset of Y, **then there exists a** compact subset C of X such that f(C) = K.

A second research-oriented subgroup incorporates topic-related bundles that denote mathematical notions, concepts or processes. Sequences in this group are built around the term *set: the set of all, be the set of, is the set of, is a set of, that the set of*. A second group of topic-related bundles revolve around the notion of *proof* and is comprised of three patterns: *the proof of theorem, in the proof of* and *(as) + the proof of the*.

3. Every ideal containing a modular ideal is itself modular. Therefore an ideal I of A is a maximal modular ideal if and only if it is maximal within **the set of all** modular proper ideals.
4. The remainder of **the proof of Theorem** 3.1 now shows that (3.2) holds for $0 < p < 2$, and the Riesz theorem gives (3.2) for $2 \leqslant p < \infty$.

Yet a third subcategory comprises 10 expressions that define concepts, describe entities or elaborate on operations and processes. Patterns such as *the definition of the, from the definition of*, and *in the definition of* all serve to discuss a concept or explain an operation whereas patterns such as *a linear combination of, is independent of the, is given by the, is of the form* and *is contained in the* are used to describe an element or operation. The expression *in such a way* is used to elaborate on an ongoing process or operation. The following statements illustrate these two functional subgroups:

5. This implies in particular that **the definition of the** Cartier operator does not depend on the choice of the separating element x.
6. What this means is that the Hamiltonian flow generated by **a linear combination of** the momentum functions consists of translations in position of the particle.
7. It is easy to verify that the isomorphism **is independent of the** choice of the metric.
8. A more general class of sub-rings than (a) **is given by the** following definition.

The fourth research-oriented group includes two bundles that serve to showcase a logical connection between two conditions. The most frequent and widely dispersed topic-related bundle in the list is the bi-directional conditional *if and only if*. A similar, though less frequent, sequence is the bundle *if there is a*, which highlights an *if–then* logical operation. Here are two examples of both types:

9. Two permutations are conjugate **if and only if** they have the same cycle structure.
10. Any imprimitive subgroup preserves a decomposition of the space as a direct sum of subspaces of the same dimension. **If there is a** form then either the subspaces are non-singular or there are precisely two of them.

A fifth research-oriented subgroup consists of location bundles that are used to indicate the place of an element or entity. Two related bundles represent this functional subgroup, namely *the right hand side* and *the left hand side* which are used to refer to the two sides of an equation.

11. Since μ divides both terms on **the left hand side**, it must divide **the right hand side**, and this completes the proof.

The final research group incorporates bundles that are utilized for quantifying elements. Quantification is a common procedure in mathematics, and is represented by four bundles: *such that for all, the sum of the, is an element of* and *a finite number of.*

12. In the other direction, suppose G is connected and suppose V is any finite-dimensional subspace of H **such that for all** $X \in g$, $V \subset \mathrm{Dom}(\pi(X))$ and $\pi(X)$ maps V into V.
13. It follows by the Residue Theorem that **the sum of the** residues of *f* on E is 0.

Turning now to text-oriented bundles, it appears that lexical bundles are used to serve several sub-functions, including framing attributes, comparing and contrasting entities, directing the reader's attention to parts of the text, announcing results and showing transition. Framing bundles include *with respect to the, in the case of, in terms of the, a special case of, an example of a* and *in this case the/we*. Due to space limitations, the following statements are examples of a selected number of bundles.

1. Remember that we enumerated the chord diagrams **with respect to the** number of crossings (= number of edges in H(C)) as highlight to Chapter 7.
2. The foregoing results can be used to prove a very useful criterion for irreducibility of certain polynomials over a function field. **A special case of** the following proposition is known as Eisenstein's Irreducibility Criterion.

Some text-oriented bundles are used to draw contrasts and comparisons between elements and entities. The comparison/contrast sub-category contains four bundles: *on the other hand, is equivalent to the, is the same a + (the)* and *is equal to the.*

3. These descriptions will be refined in Section 2 when we formally define identities. **On the other hand**, partially ordered sets and topological spaces are not readily described as universal algebras.
4. Every valuation **is equivalent to the** valuation induced by its valuation ring.

Another text-oriented sub-group consists of bundles that serve to structure the text (*in this section we, in this chapter we, we have the following, in the next section, it follows from the, it follows that the*). Here are some examples representing the two types:

5. **In this section we** study ideals of normed algebras and introduce the concept of a multiplier algebra.
6. A more subtle question is how to verify the associative law for the binary operation. This is so familiar in ordinary addition that we are prone to overlook it. When it is encountered in matrix multiplication, **it follows from the** associative law in the underlying ring.

Two text-oriented bundles are resultative markers, helping to lay out a mathematical outcome or to announce the result of a math operation. Bundles serving this function include *is said to be* and *(is) a consequence of the.*

7. The ideal I **is said to be** a maximal modular ideal if it is modular and also a maximal proper ideal.
8. This is one of the most profound results in commutative harmonic analysis, and, as usual, the proof is based on the Plancherel theorem which in turn *is* **a consequence of the** inversion formula.

The final, and by comparison the smallest, functional group incorporates bundles that perform a participant-oriented function. Lexical bundles in the first participant-oriented subgroup help to emphasize the centrality of the arguments laid out by mathematician writers. This subgroup is represented by three bundles, namely *the fact that the, from the fact that* and *and the fact that*, which revolve around the epistemic noun *fact*.

9. This is the geometric reason for **the fact that the** Burau representation can be reduced but is not a direct sum of its reduced form with a one-dimensional representation.
10. That all atoms of M belong to X follows **from the fact that** any generating subset of a monoid must contain all the atoms.A second participant-oriented group involves the use of bundles indicating the possibility or probability of carrying out a particular procedure. The probability/possibility group consists of three bundles: *can be written as, can be used to* and *can be found in*.
11. Every element of the free product with amalgamation **can be written as** a reduced word, but this representation is not unique.
12. The primary cyclic decomposition **can be used to** characterize cyclic modules via their elementary divisors.
13. Discussions of the subject under various different aspects **can be found in** the monographs by Hewitt and Ross.

Bundles forming the third group include engagement markers that are initiated by the first person plural pronoun *we*. The engagement group has three bundles: *we may assume that, we see that the* and *we say that a*. These sequences can be exemplified by the corpus-derived statements below:

14. Clearly, **we may assume that** X contains a nonconstant function.
15. In this case **we say that the** disk has rational radius.

A similar subgroup of participant-oriented bundles include incomplete clauses that are headed by the anticipatory *it* and serve as attitude markers. More specifically, they function to reflect either the ease of processing, as in the sequence *it is easy to (to see that)*, or sufficiency, as in the patterns *it is enough to* and *(it) suffices to show that*. Examples of these sequences are given below:

16. **It is easy to see** that the completeness of a normed space is invariant with respect to equivalent norms.
17. Thus **it suffices to show that** each reflection lies in the image of A.

Two final participant-oriented bundles are used to appeal to the reader. The sequence *to show that the* tells the readers about a procedure undertaken to prove a theorem. The verb *show* here is used to mean *prove*. The sequence *let (x) be a* is utilized to alert the reader to the introduction of a new series of procedures, such as proofing a theorem:

18. We leave it to the reader **to show that the** determinant function and the trace function are invariants for similarity.
19. **Let** ***x*** **be a** nonempty partially ordered set in which every nonempty chain has an upper bound.

It is important to point out here that some bundles tend to serve more than one function. Examples include patterns such as *a special case of* which may be interpreted from a stance point of view. By checking concordance lines, however, it seems obvious that this bundle is predominantly used as a framing device, thus serving a text-oriented function. The expression *if and only if* can be considered as a text-structuring bundle. After discussing this bundle with one informant, it becomes clear that this pattern is inextricably linked to the mathematical domain and would better be labeled as a research-oriented bundle.

IV Discussion

Several studies concur that ESP and mathematics students have few linguistic resources to draw on while learning mathematical language. The paucity of such resources stems from several factors, the most important of which is attributed to the complex, multi-semiotic nature of mathematics. This study is an attempt to present students, professionals and materials authors with a list of the most recurrent, widely spread and useful multiword sequences in a specialized corpus of graduate-level textbooks in the domain of mathematics. After applying a series of selection parameters and refinement procedures, the final list is composed of 65 four-word bundles, all of which have been subjected to structural and functional analyses. Although the size of the corpus upon which this study draws is comparably large, the final number of bundles is not as great as is reported in some previous studies (e.g. Hsu, 2014; Liu, 2012; Simpson-Vlach & Ellis, 2010; Wood & Appel, 2014). This may be due to the fact that this study employs a more stringent set of corpus-based and expert-informed filtering procedures. Another reason lies in the multi-semiotic nature of mathematical register that employs linguistic and non-linguistic means while constructing and disseminating knowledge (O'Halloran, 2005).

As for the grammatical forms of the bundles, it is clear that these forms vary considerably, with clausal patterns occurring in greater numbers than phrasal patterns, an outcome that goes against findings reported in some previous studies (e.g. Chen & Baker, 2010; Cortes, 2004; Pan et al., 2016). Previous research seems to suggest that mature academic writing is characteristically phrasal, whereas writing produced by less mature, less experienced writers is clausal in nature. The widespread use of clausal patterns may be interpreted from a register perspective as textbook writers appear to prefer more clause-based constructions, given that they are not under space or time constraints which compel them to adhere to a reductionist form of language (Biber et al., 2004). A second important grammatical feature of the bundles making up the list is the dominant presence of copular-be structures. Structural types as such are key for making attributive and identifying clauses (Veel, 1999). Attributive clauses are used to 'make explicit to the students the organization of uncommonsense knowledge in mathematics and play an important role in apprenticing students into mathematical knowledge', while identifying clauses are exploited to 'introduce a technical term and to negotiate between technical and less-technical construals of knowledge' (Veel, 1999, p. 195). Yet the third structural pattern emerging from the analysis is the passive-based construction. The ubiquitous presence of passive forms seems to contradict the conclusion reached by Cunningham (2017), who reported no passive verbs in her study of recurrent frames in a corpus of mathematical journal articles. This discrepancy may be explained from a genre perspective, as journal article authors are aware of 'the weakening effect of the passive voice' (Master, 1991, p. 16) as they present their own work to readers. Perhaps the impetus for obscuring agency in mathematics lies in the ever-increasing tendency among writers of this genre to value abstraction, given the inferencing nature of mathematical thinking and practice (Saitta & Zucker, 2013).

The third layer of analysis involves classifying bundles according to the discourse functions that they serve. Most bundles appear to fulfill a research-oriented function. The concentration of bundles in this group is anticipated, given that 'mathematical language is technical and often involves complex taxonomies of terms' (O'Halloran, 2005, p. 78). The analysis has uncovered some research-oriented sub-functions that have not been pinpointed in previous research. For example, some lexical patterns serve to mark the existence of a particular mathematical entity or property. Examples include bundles such as *the existence of the/a* and *then there exists a*. Other bundles marking uniqueness are also recognized, including *there exists a unique* or *there is a unique*. To account for these bundles, an existence/uniqueness sub-group is thus created. A third newly unveiled sub-group includes bundles that are used to mark conditionals. *If and only if* and *if there is a* serve to signal the implicational and causal relationships between two or more mathematical variables. Taken together, such register-specific bundles are reported in the literature under different terms, including topic-related (Hyland, 2008b), subject-specific (Jablonkai, 2010), and content-focused (Breeze, 2013). The second function served by the bundles is that of text-oriented. It seems clear that the large number of text-oriented bundles reported in this study stands in stark contrast with the number highlighted by Biber et al. (2004) whose total number of text-organizers in academic textbooks is the smallest compared with research- and participant-oriented ones. Two text-structuring sub-functions appear to have the greatest number of bundles, namely framing and structuring. Several framing and structuring sequences seem to transcend register boundaries, as they occur in other distinct disciplines (Hyland, 2008b). Examples include patterns such as *with respect to the* and *in terms of the*. There are, however, few bundles which show a strong tendency to be typical of the mathematical discourse. Mathematics-oriented bundles as such include patterns like *it follows from the* and *it follows that the*, which are used to call the reader's attention to a logical sequence of steps needed for proving, or rather disproving, a theorem.

Participant-oriented bundles make up the smallest functional group in our study. The rarity of such patterns in the data seems to be congruent with the findings reported by

Biber (2006, p. 111) and Durrant (2017). Within the participant-oriented group, there are some interesting patterns that merit some discussion. The ubiquity of the bundles initiated by the first-person plural pronoun *we* (e.g. *we may assume that* and *we see that the*) is interpreted differently by various scholars. McGrath and Kuteeva (2012) maintain that the pronoun *we* engages the reader, whereas Pimm (1984) argues that it is commonly used to establish the authorial voice of experts. An informant commenting on these results believes that the widespread use of the pronoun *we* is driven by a desire to avoid being arrogant in a domain in which knowledge is accumulatively constructed. Given Veel's (1999) emphasis that mathematical language differs markedly with respect to the registers under investigation, further research is needed to unveil the full range of uses attached to pronouns across different mathematical text types. A similar interesting tendency in the data is the large number of bundles beginning with the directive verb *let*. Graves et al. (2014, p. 6) maintains that these bundles serve to alert the reader 'for the imminent introduction of mathematical concepts including definitions and relevant notations'. The *let*-pattern appears to be very productive as it occurs throughout the data with different variables (e.g. *let m be a, let a be a, let f be a*). The analysis of participant-oriented bundles reveals a pattern of three extraposed constructions, all of which end with an infinitival verb phrase. These constructions are used to signal ease of processing, as in the bundle *it is easy to see*, or sufficiency, as in the bundles *it is enough to* and *it suffices to show*. It is clear that *it*-clauses are not as common in mathematical textbooks as in academic prose (Ädel & Erman, 2012).

V Conclusions

English learning programs are increasingly incorporating students from a cross-section of disciplinary backgrounds. While it can be easy for instructors to find materials useful for students majoring in the humanities, for example, the students aiming to pursue studies in mathematics are left at the mercy of circumstances; that is, they learn register-specific patterns if the instructor has a mathematical background or the textbook contains chapters with mathematical content. This corpus-derived expert-approved list can be readily accommodated in language programs, and mathematics students can study these patterns and identify ways in which they can be used. Items on the list are arranged based on a composite score that combines frequency of occurrence, distribution across texts and the opinions of experts. The structural and functional examples provided in this study may also be used as stand-alone statements to help exemplify mathematically-oriented bundles. Martinez and Schmitt (2012, p. 301) point out that recurrent patterns should occupy a 'prominent place in language teaching textbooks and materials, as well as texts of language achievement and proficiency'.

In conclusion, this study attempts to present a list of mathematics-oriented bundles, combining both corpus linguistics approaches and expert-informed judgment. Although the present research strives to achieve this goal, there are some limitations that need to be acknowledged. It should be noted that the list is not inclusive of all useful recurrent patterns in the mathematical register. By adhering to a predefined set of criteria for selecting and distilling data, it is possible that several key patterns did not find their way into the list simply because they did not meet the criteria for inclusion. Thus, the items on the list and their structural and functional attributes should be taken as a model to emulate while designing mathematics language pedagogy. Another limitation is that the list was compiled based on academic textbooks written for graduate-level students. Many more useful items might have been presented had the study been expanded to include other genres such as theses, dissertations and journal articles.

Language instructors can use this short list to draw attention to how mathematicians use language to advance claims and lay out arguments. Amateur mathematicians may find these lexical items of greater importance, as they endeavor to adhere to community-approved norms of expressions and by making use of these norms, they can advance mathematical arguments, construct knowledge of the subject matter, link ideas and create a persona that is both distinct and authorial.

Conflict of Interest

Also, if this study is part of a larger study or if you have used the same data in whole or in part in other papers, both already published or under review please state where the paper is published and describe clearly and in as much detail as you think necessary where the similarities and differences are and how the current manuscript makes a different and distinct contribution to the field.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The author extends his appreciation to the Deanship of Scientific Research at King Saud University for funding this work through the Research Project No R5-16-01-31.

ORCID iD

Abdullah Alasmary https://orcid.org/0000-0002-8025-5702

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*, 235–247.

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*, 81–92.

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, *5*, 97–116.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*, 263–286.

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

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

Breeze, R. (2013). Lexical bundles across four legal genres. *International Journal of Corpus Linguistics*, *18*, 229–253.

Brilliant-Mills, H. (1994). Becoming a mathematician: Building a situated definition of mathematics. *Linguistics and Education*, *5*, 301–334.

Chan, S. (2015). Linguistic challenges in the mathematical register for EFL learners: Linguistic and multimodal strategies to help learners tackle mathematics word problems. *International Journal of Bilingual Education and Bilingualism*, *18*, 306–318.

Chen, Y-H., & Baker, P. (2010). Lexical Bundles in L1 and L2 Academic Writing. *Language Learning & Technology*, *14*, 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.

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

Cunningham, K.J. (2017). A phraseological exploration of recent mathematics research articles through key phrase frames. *Journal of English for Academic Purposes*, *25*, 71–83.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, *28*, 157–169.

Durrant, P. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics*, *38*, 165–193.

Grabowski, L. (2015). Keywords and lexical bundles within English pharmaceutical discourse: A corpus-driven description. *English for Specific Purposes*, *38*, 23–33.

Graves, H., Moghaddasi, S., & Hashim, A. (2014). 'Let G = (V, E) be a graph': Turning the abstract into the tangible in introductions in mathematics research articles. *English for Specific Purposes*, *36*, 1–11.

Herbel-Eisenmann, B., Wagner, D., & Cortes, V. (2010). Lexical bundle analysis in mathematics classroom discourse: The significance of stance. *Educational Studies in Mathematics*, *75*, 23–42.

Hsu, W. (2014). The most frequent opaque formulaic sequences in English-medium college textbooks. *System*, *47*, 146–161.

Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*, 41–62.

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

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, *41*, 235–253.

Jablonkai, R. (2010). English in the context of European integration: A corpus-driven analysis of lexical bundles in English EU documents. *English for Specific Purposes*, *29*, 253–267.

Kuteeva, M., & McGrath, L. (2015). The Theoretical Research Article as a Reflection of Disciplinary Practices: The Case of Pure Mathematics. *Applied Linguistics*, *36*, 215–235.

Leung, C. (2005). Mathematical vocabulary: Fixers of knowledge or points of exploration? *Language and Education*, *19*, 126–134.

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes*, *31*, 25–35.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, *33*, 299–320.

Master, P. (1991). Active verbs with inanimate subjects in scientific prose. *English for Specific Purposes*, *10*, 15–33.

McGrath, L. (2016). Open-access writing: An investigation into the online drafting and revision of a research article in pure mathematics. *English for Specific Purposes*, *43*, 25–36.

McGrath, L., & Kuteeva, M. (2012). Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. *English for Specific Purposes*, *31*, 161–173.

Morgan, C. (2005). Words, Definitions and Concepts in Discourses of Mathematics, Teaching and Learning. *Language and Education*, *19*, 103–117.

O'Halloran, K. (1998). Classroom Discourse in Mathematics: A Multisemiotic Analysis. *Linguistics and Education*, *10*, 359–388.

O'Halloran, K. (2005). *Mathematical discourse: Language, symbolism and visual images*. London: Continuum.

O'Halloran, K. (2015). The language of learning mathematics: A multimodal perspective. *Journal of Mathematical Behavior*, *40*, 63–74.

Pan, F., Reppen, R., & Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals. *Journal of English for Academic Purposes*, *21*, 60–71.

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.

Pimm, D. (1984). Who is we?. *Mathematics Teaching*, *107*, 39–42.

Qin, J. (2014). Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics. *System*, *42*, 220–231.

Saitta, L., & Zucker, J. (2013). *Abstraction in artificial intelligence and complex systems*. New York: Springer.

Schleppegrell, M. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading and Writing Quarterly*, *23*, 139–159.

Scott, M. (2016). *WordSmith Tools version 7* [software]. Stroud: Lexical Analysis Software.

Shin, D., & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal*, *62*, 339–348.

Simpson-Vlach, R., & Ellis, N.C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*, 487–512.

Veel, R. (1999). Language, knowledge and authority in school mathematics. In F. Christie, (Ed.), *Pedagogy and the shaping of conciousness: Linguistics and social processes* (pp. 185–216). London: Continuum.

Wood, D.C., & Appel, R. (2014). Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes*, *15*, 1–13.

Appendix 1

List of books used for the corpus analysis of academic lexical bundles

1. Aigner, M. (2007). *A course in enumeration*. Volume 238. Berlin: Springer Science & Business Media.
2. Albiac, F., & Kalton, N.J. (2006). *Topics in Banach space theory*. New York: Springer.
3. Applebaum, D., & Heyer, H. (2014). *Probability on compact Lie groups*. Cham: Springer.
4. Clarke, F.H. (2013). *Functional analysis, calculus of variations and optimal control*. London: Springer.
5. Cohen, M.M. (2012). *A course in simple-homotopy theory*. Volume 10. Berlin: Springer Science & Business Media.
6. Conforti, M., Cornuéjols, G., & Zambelli, G. (2014). *Integer Programming*. Cham: Springer.

7. Diamond, F., & Shurman, J.M. (2005). *A first course in modular forms*. Volume 228. New York: Springer.
8. Eisenbud, D. (2005). *The geometry of syzygies: A second course in commutative algebra and algebraic geometry*. New York: Springer.
9. Everest, G., & Ward, T. (2006). *An introduction to number theory*. Volume 232. Berlin: Springer Science & Business Media.
10. Garnett, J.B. (2007). *Bounded Analytic Functions*. New York: Springer.
11. Geoghegan, R. (2010). *Topological methods in group theory*. New York: Springer.
12. Gilman, J.P., Kra, I., & Rodríguez, R.E. (2013). *Complex analysis: In the spirit of Lipman Bers*. New York: Springer.
13. Goodman, R.W., & Wallach, N.R. (2009). *Symmetry, representations and invariants*. New York: Springer.
14. Grafakos, L. (2014). *Classical Fourier analysis*. New York: Springer.
15. Grillet, P.A. (2011). *Abstract algebra*. New York: Springer.
16. Grubb, G. (2010). *Distributions and operators*. New York: Springer.
17. Hall, B.C. (2013). *Quantum theory for mathematicians*. New York: Springer.
18. Hartshorne, R. (2010). *Deformation theory*. New York: Springer.
19. Hilton, P.J., & Stammbach, U. (2012). *A course in homological algebra*. Volume 4. Berlin: Springer Science & Business Media.
20. Jorgensen, P.E. (2006). *Analysis and probability: wavelets, signals, fractals*. Volume 234. Berlin: Springer Science & Business Media.
21. Kaniuth, E. (2008). *A course in commutative Banach algebras*. Volume 246. Berlin: Springer Science & Business Media.
22. Kassel, C., & Turaev, V.G. (2011). *Braid groups*. New York: Springer.
23. Krantz, S.G. (2013). *Geometric analysis of the Bergman kernel and metric*. New York: Springer.
24. Lee, J.M. (2013). *Introduction to smooth manifolds*. New York: Springer.
25. MacCluer, B.D. (2010). *Elementary functional analysis*. New York: Springer.
26. Penot, J.-P. (2013). *Calculus without derivatives*. New York: Springer.
27. Petersen, P. (1998). *Riemannian geometry*. New York: Springer.
28. Ratcliffe, J.G. (2011). *Foundations of hyperbolic manifolds*. New York: Springer.
29. Roman, S., Axler, S., & Gehring, F.W. (2005). *Advanced linear algebra*. Volume 3. New York: Springer.
30. Serre, D. (2011). *Matrices: Theory and applications*. New York: Springer.
31. Silverman, J.H. (2007). *The arithmetic of dynamical systems*. New York: Springer.
32. Stichtenoth, H. (2009). *Algebraic function fields and codes*. Volume 254. Berlin: Springer Science & Business Media.
33. Stroock, D.W. (2011). *Essentials of integration theory for analysis*. New York: Springer.
34. Wells, R.O. (2011). *Differential analysis on complex manifolds*. New York: Springer.
35. Wilson, R.A. (2009). *The finite simple groups*. London: Springer.
36. Zhu, K. (2012). *Analysis on Fock Spaces*. New York: Springer.

Appendix 2. Final list of lexical bundles.

Number	Sequences	Functional category/ sub-function	Frequency (pmw)	Distribution (%)	Experts	T score
1.	if and only if + (the)(it is)(there)	Research/ conditional	911	100	100	89.88
2.	that there is a/an	Research/existence	128	100	88	57.53
3.	the set of all	Research/ topic-related	133	94.44	91	57.12
4.	(as) + the proof of the	Research/ topic-related	99	97.22	90	56.65
5.	we may assume that	Participant/stance	138	94.44	90	56.64
6.	in the proof of	Research/ topic-related	128	100	86	56.24
7.	there is a unique	Research/uniqueness	98	91.67	93	56.07
8.	is said to be	Text/resultative	130	91.67	90	55.14
9.	to show that the	participant/reader engagement	68	94.44	91	55.08
10.	with respect to the	text/framing	107	97.22	87	54.98
11.	if there is a	Research/conditional	54	94.44	91	54.64
12.	then there exists a	Research/ existence	65	86.11	96	54.47
13.	is equivalent to the	Text/contrast & comparison	48	94.44	91	54.45
14.	show that there is	Research/existence	48	97.22	89	54.41
15.	be the set of	Research/ topic-related	76	94.44	89	54.04
16.	the definition of the	Research-oriented/ definitional	39	94.44	90	53.52
17.	on the other hand	Text/contrast & comparison	159	100	80	53.35
18.	in the definition of	Research/definitional	26	88.89	94	53.20
19.	it follows from the	Text/structuring	45	88.89	93	53.15
20.	is contained in the	Research/description	41	91.67	91	52.99
21.	then there is a/an	Research/existence	97	91.67	88	52.82
22.	is the set of	Research/ topic-related	71	94.44	87	52.60
23.	that there exists a	Research/existence	54	86.11	93	52.19

(Continued)

Appendix 2. (Continued)

Number	Sequences	Functional category/ sub-function	Frequency (pmw)	Distribution (%)	Experts	T score
24.	the proof of theorem	Research/ topic-related	131	86.11	88	51.40
25.	in this section we	Text/structuring	92	94.44	84	51.33
26.	it is easy to + (to see that)	Participant/ease of processing	133	97.22	80	51.29
27.	(it) + suffices to show that	Participant/sufficiency	63	83.33	93	51.23
28.	let (*x*)* be a	Participant/reader engagement	118	77.78	94	51.12
29.	the existence of the/a	Research/existence	33	88.89	90	50.85
30.	a linear combination of	Research/description	30	86.11	92	50.79
31.	a special case of	Text/framing	28	86.11	92	50.73
32.	is independent of the	Research/description	27	91.67	88	50.62
33.	in the case of	Text/framing	66	91.67	86	50.56
34.	a finite number of	Research/ quantification	34	83.33	93	50.32
35.	from the definition of	Research/definitional	35	86.11	91	50.31
36.	is given by the	Research/procedure	33	91.67	87	50.16
37.	it follows that the	Text/structuring	45	88.89	88	49.94
38.	the fact that the	Participant/centrality	47	97.22	82	49.87
39.	is an element of	Research/ quantification	26	80.56	94	49.47
40.	is equal to the	Text/contrast & comparison	38	80.56	93	49.20
41.	in terms of the	Text/framing	61	97.22	80	49.03
42.	an example of a	Text/framing	30	86.11	89	48.86
43.	is the same as + (the)	Text/contrast & comparison	45	94.44	82	48.56
44.	such that for all	Research/ quantification	56	75	95	48.56
45.	there exists a unique	Research/uniqueness	41	75	95	48.09
46.	in this case the/ we	Text/framing	27	88.89	86	48.08
47.	the right hand side	Research/location	68	91.67	82	48.05

(Continued)

Appendix 2. (Continued)

Number	Sequences	Functional category/ sub-function	Frequency (pmw)	Distribution (%)	Experts	T score
48.	the sum of the	Research/ quantification	31	77.78	93	47.74
49.	we say that a	Participant/ engagement	29	80.56	91	47.63
50.	can be used to	Participant/possibility	27	86.11	87	47.48
51.	from the fact that	Participant/centrality	36	88.89	84	47.08
52.	can be written as	Participant/ probability	47	88.89	83	46.78
53.	is a set of	Research/ topic-related	33	77.78	90	45.87
54.	(is) + a consequence of the	Text/resultative	29	77.78	90	45.74
55.	in the next section	Text/structuring	31	86.11	84	45.68
56.	we have the following	text/structuring	35	80.56	87	45.25
57.	in this chapter we	Text/structuring	38	88.89	81	45.21
58.	it is enough to	Participant/sufficiency	30	75	91	45.17
59.	the left hand side	Research/location	32	80.56	87	45.15
60.	and the fact that	Participant/centrality	30	80.56	86	44.45
61.	we see that the	Participant/ engagement	46	80.56	85	44.31
62.	that the set of	Research/ topic-related	27	75	89	43.79
63.	can be found in	Participant/ probability	25	75	85	41.16
64.	is of the form	Research/description	36	75	84	40.86
65.	in such a way	Research/elaboration	26	80.56	80	40.46

Notes. * other frames with different variables (*x*) also exist. PMW = Per Million Words.