문학석사 학위논문

A Corpus-based Study of Lexical Bundles in Four Academic Disciplines

4개 학문분야에서의 어휘다발에 대한 코퍼스 기반 연구

2020년 2월

서울대학교 대학원

영어영문학과 영어학 전공

최 봉 준

Abstract

A Corpus-based Study of Lexical Bundles in Four Academic Disciplines

Bongjun Choi

Department of English Language and Literature

Seoul National University

The use of lexical bundles in four different academic disciplines is studied in this thesis. The disciplines in interest for a multidisciplinary analysis are applied linguistics, management, engineering and microbiology because they are sought to appropriately reflect the characteristics of the academic fields of arts and humanities, social sciences, physical sciences and life sciences, respectively. At least 500,000 words were collected per discipline to make up a corpus of approximately 2.5 million words. Research articles from renowned international academic journals in each discipline were collected to build the corpus for this study. Three features are mainly analyzed: 1) frequency comparison of lexical bundles, 2) structural comparison of lexical bundles following Biber *et al.*'s (1999) taxonomy, and 3) functional comparison of lexical bundles based on Hyland's (2008b) categorization. It was found that engineering outnumbered all other disciplines in both types (number of distinct bundles) and tokens (number of total bundles). Additionally, lexical bundles were generally used uniquely within discipline, with 70% of all lexical bundles in engineering not occurring in any other discipline. In terms of structural characteristics, disciplines in the hard knowledge fields (engineering and microbiology) significantly used more lexical bundles of passive structure than disciplines in the soft knowledge fields (applied linguistics and management). Researchers' reluctance to show authorial stance in their writing was closely related to this tendency. Authors in engineering and microbiology preferred hiding it while authors in management were less reluctant to showing stance. Functionally, it was interesting that text-oriented lexical bundles were used the most in all four disciplines. The reason for such predominance of text-oriented bundles was hypothetically explained to be due to some characteristics of research articles, as opposed to student writing. As such, by studying how lexical bundles are used similarly or differently across different academic disciplines, this research is expected to pedagogically aid in discovering disciplinary specificity since it sheds light to prevalent, formulaic lexical conventions of each disciplinary community.

Keywords : Lexical bundle, corpus-based study, structure, function, disciplinary variation

Student Number : 2018-28156

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1. Motivation and background

You shall judge a word by the company it keeps (Firth, 1951). As Firth's famous slogan suggests, there are numerous instances in which a word is meaningful and can be interpreted only in relation to the words adjacent to it. Supporting this argument is Altenberg's (1998) research in which he estimated that recurrent word combinations make up 80% of the words in the London-Lund corpus. Although less dramatic, Erman & Warren (2000) similarly examined the Lancaster-Oslo-Bergen corpus and revealed that 52.3% of the written corpus were prefabricated word combinations, which they called *prefabs*. Acknowledging such pervasive collocational nature of language, researchers in various subfields of applied linguistics went on to study whether such formulaicity has an influence on language both theoretically and empirically. Theoretically, Sinclair (1991) proposed a new, radical way of looking at grammar, arguing that "most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up a text. This is totally obscured by the procedures of conventional grammar (p. 108)". Similarly, Hoey (2005) observed sentences comprising bundles that are interlocked as individual words are *primed* mentally for operation with other words based on our experience of those words in frequent encounter. Sinclair and Hoey did not see lexical choices traditionally as being constrained by separate grammatical slots, but as systematically structured via frequent patterns of use. Empirically, most research focused on evaluating whether there is a relationship between the nativity or proficiency and the use of formulaic language, that is, if native speakers or highly proficient speakers are more likely to have competence and are able to flexibly and fluently employ these word combinations in speech or writing compared to non-native or novice speakers or writers. Regarding this, the general consensus is that there is a positive relationship between the nativity or proficiency and the use of recurrent word constructions, as Haswell (1991) stated, “there can be little doubt that as writers mature they rely more and more on collocations and that the lesser use of them accounts for some characteristic behavior of apprentice writers (p. 236)”. McCulley (1985) similarly pointed out that the use of fixed expressions and collocations can be considered proficient use of language.

Nattinger & DeCarrico (1992) also emphasized the importance of frequent multi-word sequences to be a means of aiding communication, as they shape the language into a more predictable form to the hearer. Cortes (2004) also described that frequently using fixed expressions “seems to signal competent language use within a register to the point that learning conventions of register use may in part consist in learning how to use certain fixed phrases (p. 398)”. Furthermore, the usefulness of lexical bundles is well-summarized in Coxhead & Byrd (2007) to indicate that:

(a) for advanced students, writing tasks become easier because they can construct sentences in larger ready-made chunks as opposed to having to construct sentences word-by-word,

(b) because lexical bundles are frequent in academic writing, they offer students a clear indicator of fluent writing practices and reader expectations in academia, and

(c) lexical bundles allow students to see the different elements present in the interaction between syntax and lexis.

Essentially, being able to interpret and smoothly construct academic texts using lexical bundles can be a sign of a proficient writer or reader. Owing to such pedagogically advantageous characteristics of formulaicity, pedagogy-oriented studies have flourished in the literature. Teaching formulaic language to non-native learners of language, however, is not an easy task even to the highly proficient non-native speakers, as discussed in detail in Wray's (2000) argument.

> From the bizarre idiom, through the customary collocation, to the turns of phrase that have no other apparent linguistic merit than that 'we just say it that way', the subtleties of a language may floor even the proficient non-native, not so much because of a non-

> alignment between interlanguage and target language forms, as because the learner lacks the necessary sensitivity and experience that will lead him or her unerringly away from all the grammatical ways of expressing a particular idea except the most idiomatic. (p. 463)

Nevertheless, there have been constant endeavors to discover the ways in which teaching formulaic language would maximally benefit language learners. Efforts of this kind include studies that attempt to identify how native and non-native speakers use lexical bundles differently by contrastively analyzing the overused and underused lexical bundles in terms of grammatical structure and function of the bundles (Chen & Baker, 2010; Adel & Erman, 2012; Salazar, 2010; Choi, 2015). Furthermore, some research aimed at unearthing the efficient ways to explicitly teach the proper use of lexical bundles in the classroom. For example, Cortes (2006) conducted a pre- and post-instruction analysis of lexical bundles and found that while there was no difference in the bundle production, students' awareness increased in post-instruction. Replicating this study, Kim (2018) found that there was an increase in the use of lexical bundles in post-instruction, although not significant considering the time spent in training.

Along with the native, non-native distinction, there has also been evidence of variation depending on context or format of language. For instance, Biber (2006) found that the number of different bundles used in spoken, classroom environment was approximately double than that of casual conversation data and nearly four times than that of written, textbook language. In addition, Cortes (2004) revealed in her study using biology and history papers as data that the use of lexical bundles in university students' writings were significantly dissimilar to the use of bundles that are typical in published academic writing. Modern academia is constantly expanding and evolving, giving birth to new research types such as the stand-alone literature reviews. These novel text types are referred by Noguchi (2006) as a *bridge genre*, helping an individual to become part of a particular discourse community. Wright (2019) studied how lexical bundles are used differently in these stand-alone literature reviews. Based on her 3.4 million-word corpus of stand-alone literature reviews in education, psychology and medicine, she found similarities in the use of lexical bundles and suggested that a core set of bundles for written academic prose may exist. She further argued that the function of these bundles was to establish research gaps, clarify results and methods, and report data. Similarly, Sisilia *et al*. (2019) looked into the use of lexical bundles in the literature review sections of undergraduate students' final projects and found that noun phrase-based bundles and research-oriented bundles were most commonly used. Also, Jalali & Moini (2014) studied the structure of lexical bundles used in the introduction section of research articles within the academic discipline of medicine. Jalali & Moini (2014) showed that most researchers in medicine used noun phrase and phrasal bundles in the introduction section to establish their academic discourse.

This thesis takes a different viewpoint and attempts to add to the relatively lacking literature, focused on similarities and differences of lexical bundle usage across academic disciplines. The disciplines in interest for a multidisciplinary analysis are applied linguistics, management, engineering, and microbiology because they are sought to appropriately reflect the characteristics of the academic fields of arts and humanities, social sciences, physical sciences and life sciences, respectively.

1.2. Research questions

The aim of this study is mainly to discover similarities and peculiarities of lexical bundles used across the four academic disciplines so as to contribute to language learning in the classroom environment. General lexical tendencies could aid general English course instructors, for example, and lexical peculiarities within each discipline could be advantageous since it provides insight to the convincing lexical conventions that professionals use frequently in the respective academic discipline. This study thus attempts to answer the following questions.

1. What are the similarities and differences in the use of lexical bundles across the four disciplines (applied linguistics, management, engineering, microbiology) in terms of frequency?

2. What are the structural similarities and differences in the use of lexical bundles across the four disciplines (applied linguistics, management, engineering, microbiology)?

3. What are the functional similarities and differences in the use of lexical bundles across the four disciplines (applied linguistics, management, engineering, microbiology)?

1.3. Organization of the study

The next chapter defines and classifies lexical bundles and the characteristics that define them. Studies that attempt to unearth the disciplinary variation in the use of lexical bundles will also be summarized in detail. Mainly, Hyland's (2008b) research will be covered because it is the basis of the functional classification of lexical bundles in this thesis. Although not in the context of disciplinary variation, Biber *et al.*'s (1999) genre-focused research will also be covered since the current study's structural classification of lexical bundles is based on this research. Chapter 3 first describes the characteristics of the data used in this study, namely from which journals the research articles were collected and how the structural development of each research article was matched. Next, significant issues of data collecting such as manual exclusion and the tools will be discussed. The structural and functional taxonomy that is adopted for the present study is then specifically addressed. Chapter 4 elaborates on the data analysis and results of the different and similar lexical bundle usages across disciplines, starting with the frequency comparison. The number of total bundles and distinct bundles are presented along with an overview of the top 20 bundles in each discipline. Next, the structural classification of lexical bundles is addressed mainly in terms of part-of-speech groupings: noun phrase-based, prepositional phrase-based and verb phrase-based lexical bundles. Wrapping up Chapter 4, the analysis of the functional characteristics of lexical bundle use, mainly based on whether the bundles are oriented in research, text or participants, is discussed. Chapter 5 overviews and summarizes the findings of this research and discusses some implications and limitations.

Chapter 2. Previous Literature

2.1. Lexical bundles

Various terminology has been used so far to refer to these recurrent word combinations. Albeit its growing importance, it is surprising that there is no consensus on the characteristics that define them and the methodologies in identifying them. Also, there is little consensus on even what they should be called. They are referred to as *clusters* (Scott, 2006), *n-grams* (Stubbs & Barth, 2003), *formulaic sequences* (Wray, 2002), *recurrent word combinations* (DeCock, 1998), *multi-word expressions* (Siyanova-Chanturia & Martinez 2015), *lexical chunks* (O'Keeffe *et al.*, 2007), *lexical phrases* (Nattinger & DeCarrico, 1992), *lexical bundles* (Biber *et al.*, 1999), and so on. Refer to Wray (2002) for a comprehensive list of 50 different terms. For the purposes of this study, this thesis will hereafter adopt Biber *et al.*'s (1999) terminology and refer to them as *lexical bundles*. In Biber *et al*. (1999), lexical bundles are defined as "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" and "bundles of words that show a statistical tendency to co-occur (p. 990)". Chen & Baker (2010) similarly defines them as "fixed multi-word units that have customary pragmatic and/or discourse functions, used and recognized by the speakers of a language within certain contexts (p.30)."

Three defining characteristics of lexical bundles which distinguish them from other sorts of formulaic expressions are presented in Biber & Barbieri (2007). First, lexical bundles are extremely common. This characteristic raises the question "how common?" which is an important criterion (cut-off frequency) in identifying lexical bundles. This criterion is adopted among researchers in a somewhat arbitrary fashion, ranging from the lenient 10 occurrences per million words to the conservative 40 occurrences per million words. Solely relying on this frequency criterion in identifying lexical bundles could be dangerous because the idiosyncratically frequent use of lexical bundles by one or few authors could also be identified as target bundles. Although used frequently, the bundles thus overused are hard to see as bundles that represent the respective disciplines since they are author specific. To prevent such bundles from appearing in the target bundle list, the second criterion, dispersion, is also considered in identifying lexical bundles. This criterion constrains how widely a bundle has to be used across authors in order for them to be identified as target bundles to guarantee the disciplinary generality. Similar to the cut-off frequency criterion, the dispersion criterion is also adopted in an arbitrary manner. The details of these two criteria will be covered wholly in Section 3.2.2. Second, most lexical bundles are not idiomatic in meaning and not perceptually salient, whereas idioms are. For example, the meaning of the idiom, *kick the bucket* is created newly as a whole, not by combining the meaning of each and every word.

In other words, it is idiomatic in meaning and perceptually salient, while the meaning of lexical bundles such as *at the same time* is derivable by the sum of all individual words' meaning. Third, lexical bundles usually do not represent a complete structural unit, demonstrating in Biber *et al*. (1999) that most bundles in academic prose are phrasal and usually bridge two different phrases.

2.2. Lexical bundles and disciplinary variation

As mentioned above, studying lexical bundles to pedagogically aid second language learners and non-native speakers in classroom environment has prospered in applied linguistics (Chen & Baker, 2010; Adel & Erman, 2012; Salazar, 2010; Choi, 2015). Discovering genre-specific peculiarities in using lexical bundles, such as the spoken, written distinction or the academic, non-academic distinction was also a topic of interest for researchers (Biber *et* al., 1999; Biber & Barbieri, 2007; Cortes, 2004; Hyland, 2008a). However, there are not many resources in the literature that an apprentice writer in a specific discipline can reach out in order to learn the formulaic linguistic practices and conventions that members within their respective discipline find credible. Such discipline-focused approaches generally diverge into either being interdisciplinary (studying and comparing multiple disciplines) or intradisciplinary (studying a single discipline in depth). In the following sections, some interdisciplinary and intradisciplinary studies will be introduced after two main studies that were the basis for this study are introduced.

2.2.1. Hyland (2008)

One interdisciplinary study is that of Hyland (2008b), in which he compared the use of lexical bundles across the fields of electrical engineering, biology, business studies, and applied linguistics. It could be said that the current study is based on Hyland's (2008b) framework, in a way replicating the premise of his study since the functional classification of lexical bundles is based on his taxonomy. Additionally, the four disciplines that are studied in the present study were selected following Hyland's (2008b) research.

To first explain the data used in his research, the corpus was made up of research articles, doctoral dissertations, and master theses in order to capture any identifiable differences among not only academic disciplines, but also the three text types (or author level). Table 1 below summarizes Hyland's (2008b) corpus makeup. The current study consulted the four disciplines Hyland (2008b) selected since they seem to appropriately reflect the characteristics of the general fields of academic study or "a cross section of academic practice (p. 8)" (arts and humanities, social sciences, physical sciences, life sciences).

Table 1. Corpus makeup in Hyland (2008b)

Discipline	Articles	Doctoral	Masters	Totals
Electrical engineering	107,700	334,800	190,000	632,500
Biology	143,500	458,000	192,600	794,100
Business studies	214,900	437,200	192,300	844,400
Applied linguistics	211,400	670,000	248,000	1,129,400
Totals	677,500	1,900,000	822,900	3,400,400

Based on the numerous acknowledged studies that are being published in the literature, it could be said that a reasonably adequate size to conduct corpus-based analysis was collected per subcorpora, totaling 3.4 million words. As can be seen in the table above, corpus size among discipline and among text type (level) is not matched in terms of absolute word count; doctoral dissertations were collected far more than any other text type and the corpus size of applied linguistics is almost double than that of the electrical engineering corpus. However, this was not a problem since there was an appropriate normalization process in the overall research procedure to ensure the comparability across discipline and text type. 4-word bundles were the target of analysis for they are more common than 5-word bundles and have a clearer display of structures and functions compared to 3-word bundles.

Looking into the frequency, structure, and function of lexical bundles in the data explained above, he revealed that writers in each field rely on different resources in developing their arguments, establishing their reliability and persuading the audience, evidenced by the unique distribution of lexical bundles in each discipline (over 50% of lexical bundles in each discipline does not appear in other disciplines). Specifically, first in terms of frequency of the 4-word lexical bundles, he found disciplinary differences such as electrical engineering showing the widest range of bundles (213 different bundles) and biology the narrowest (131 different bundles). Structural differences were also found. The most representative of these was possibly passive structures. Passive structures were employed significantly more in the science and engineering texts than in the soft knowledge fields (applied linguistics and business studies). Functionally, research-oriented bundles were used significantly more in the hard knowledge fields (electrical engineering and biology) and text-oriented bundles were predominant in the soft knowledge fields. One of the few general tendencies observed commonly in all four disciplines was that most bundles were fragments of preposition or noun phrases.

Along with discovering such similarities and differences, his research is also noteworthy for proposing a new taxonomy of the function of lexical bundles by modifying Biber *et al.*'s (2004) taxonomy. He states that modification was necessary due to the different composition of the corpus; Biber *et al.*'s (2004) corpus was composed of multiple genres (textbooks, institutional texts, casual conversation, and so on) while Hyland's (2008b) focus was exclusively on academic writing. His framework is organized according to three broad functional orientations (research-oriented, text-oriented, participant-oriented) with more specific roles indicating more specific sub-categories. The taxonomy, which this study also adopts by further slightly modifying it, is presented in Table 2.

Table 2. Hyland's (2008b) functional taxonomy of lexical bundles

Function	**Description (Examples)**
Research-oriented	Helps writers to structure their activities and experiences of the real world
Location	Indicate time/place (*at the beginning of, in the present study*)
Procedure	Indicate events, actions and methods (*the use of the, the role of the*)
Quantification	Indicate measures, quantities, proportions and changes thereof (*a wide range of, the magnitude of the*)
Description	Indicate quality, degree and existence (*the structure of the, the size of the*)
Topic	Related to the field of research (*in the Hong Kong, the currency board system)*
Text-oriented	Concerned with the organization of the text and its meaning as a message or argument
Transition signals	Establishing additive or contrastive links between elements (*on the other hand, in addition to the)*
Resultative signals	Mark inferential or causative relations between elements (*as a result of, it was found that*)
Structuring	Text-reflexive markers which organize stretches of discourse or direct reader elsewhere in the text (*in the present study*)
Framing	Situate arguments by specifying and limiting conditions (*in the case of, with respect to the*)

Participant-oriented	These are focused on the writer or reader of the text
Stance	Convey the writer’s attitudes and evaluations (*are likely to be, may be due to*)
Engagement	Address readers directly (*it should be noted that, as can be seen*)

To briefly add to Hyland’s description of functions in the table above, research-oriented lexical bundles are usually related to the research (usually experiments) itself, rather than the interpretations and analyses that come after it. Text-oriented lexical bundles generally deliver propositions or logical arguments, frequently by referring to relationships. In a research article, for example, the author’s analyses, interpretations, implications and conclusions usually abound with text-oriented lexical bundles whereas research-oriented lexical bundles are likely to be used frequently in the methodology section. The role that participant-oriented bundles play seems quite straightforward. They are concerned directly with either the author or the reader of the text.

2.2.2. Biber *et al*. (1999)

Although not in the context of disciplinary variation, the monumental work of Biber *et al*. (1999) is also important in discussing the structural classification of lexical bundles. The current study adopts their classification without modification because it is considered well-established and reliable in the literature, proven by the fact that majority of researchers adopt it when studying the structure of lexical bundles.

In Biber *et al.* (1999) broadly discussing grammar in general, a small chapter (p.987-1036) is devoted to describing lexical expressions (bundles) in speech and writing. In their description of lexical bundles in the written genre, they use their academic subcorpora of over 5 million words to retrieve a comprehensive list of 4-word, 5-word and 6-word lexical bundles defined as word combinations that occur at least ten times per million words. As briefly mentioned above in Section 2.1., this definition criterion is considered the most lenient in the literature, inclusively identifying almost every instance of lexical co-occurrence as lexical bundles. Even relatively infrequent word sequences that may have appeared adjacently by chance are likely to be identified as lexical bundles, so substantial manual exclusion is speculated to have been done in their work to distill down their draft list into a target list to be analyzed.

The main purpose of their research was to discover some peculiarities or similarities between the spoken and written genre, but only the academic portion of their corpus is referred here since it is relevant to this study. The academic portion of their corpus includes book extracts and research articles. The total number of words in book extracts and research articles is 2,655,000 and 2,676,800, respectively. Most of the books were trade books written for specialized authorship and required background technical knowledge from the respective disciplines. The research article portion was made up of either journal articles or papers published in an edited collection from various disciplines (13 disciplines including agriculture, biology/ecology, chemistry, computing, etc.).

Regarding their data, a potential limitation could be the unmatched number of texts and number of words in the research article extracts across disciplines. For example, a single author could hypothetically represent the entire field of computing while various authors are melted together to represent the field of medicine. Moreover, one extensive article of 39,400 words was collected in computing while 257 short articles averaging only of 3,000 words were collected in medicine. This seemed not to be an issue since they did not focus on the disciplinary differences, but rather compared the whole academic subcorpora to other registers. Discipline-specific studies like the present one must ensure that there are no such large discrepancies across disciplines in order to obtain comparability and generalizability of the findings.

Upon inspection of the retrieved list based on the above academic corpus, Biber *et al*. (1999) observed that lexical bundles have clear structural correlates which makes it possible to classify them according to certain fundamental structural types. The resulting 12 classifications of correlates is presented in Table 3.

Table 3. Structural classification of lexical bundles (Biber *et al.*, 1999)

Structure	**Example**
Noun phrase with *of*-phrase fragment (NP + *of*)	*the nature of the,* *the effect of the*
Noun phrase with other post-modifier fragments (NP)	*the fact that the,* *the degree to which*
Pronoun/noun phrase + *be* (+ …)	*this is not the,* *there was no significant*
Prepositional phrase with embedded *of*-phrase fragment (PP + *of*)	*in the case of,* *in terms of the*
Other prepositional phrase (fragment) (PP)	*on the other hand,* *at the same time*
Anticipatory *it* + verb phrase/adjective phrase (Anticipatory *it*)	*it is possible that,* *it is important to*
Passive verb + prepositional phrase fragment (Passive verb)	*is shown in fig,* *was used as the*
Copula *be* + noun/adjective phrase (Copula *be*)	*is due to the,* *is consistent with the*
(Verb phrase +) *that*-clause fragment ((V) + *that*)	*these results suggest that,* *that the relationship between*
(Verb/adjective +) *to*-clause fragment ((V/A) + *to*)	*are more likely to,* *is negatively related to*
Adverbial clause fragment (Adv. cl.)	*as shown in fig,* *as can be seen*
Other expressions (Others)	*as well as the, play a role in*

As can be seen in the table, lexical bundles including noun phrases and prepositional phrases are basically classified binarily into having and lacking “of”. Also, it was mentioned in Section 2.1. above as a defining characteristic that lexical bundles do not represent complete structural units, and this is well displayed in the table above. Most of the examples are fragments of a bigger, complete structural unit and usually function to bridge two phrases.

An interesting finding in their study was that these structural correlates differed in spoken and written register; most bundles used in conversation are clausal, for example the pronoun + verb + complement construction (e.g. *I want you to*), while bundles are mostly phrasal as being parts of noun phrases or prepositional phrases (e.g. *on the other hand*), for instance, in academic prose. Further supporting the formulaic nature of language discussed at the beginning of Chapter 1, they also revealed that lexical bundles are extremely common in both conversation and academic prose. Specifically, 3-word bundles occurred more than 80,000 times per million words in conversation; over 60,000 times per million words in academic prose and 4-word bundles over 8,500 times per million words in conversation, and 5,000 times per million words in academic prose.

2.2.3. Other multidisciplinary studies

Along with Hyland (2008b), another interdisciplinary study is Durrant's (2017), which similarly studied the differing characteristics of lexical bundle use across different disciplines but took a different approach in categorizing the disciplinary groupings, pointing out to possible problematic issues that may arise from uncritically embracing the conventional grouping. Specifically, rather than assuming the disciplinary categories at the beginning of the analysis, he let the categories to first rise from a pre-analysis. Disciplinary groupings such as humanities, social sciences thus emerged based on the co-occurring lexical similarities of the pre-analysis. He then characterized the use of lexical bundles in the emerged humanities and social science (soft sciences) writings as having "a focus on abstract constructs" and "emphasizing the role of unique autonomous agents in processes that are difficult to control", while science and technology (hard sciences) writings focused "on the physical world" and emphasized "the role of passive, interchangeable, instruments in processes that are tightly controlled by the researcher (p. 190)".

Gilmore & Miller (2018), on the other hand, took an intradisciplinary approach and looked into a single discipline (civil engineering) to discover disciplinary specificity for a restricted target audience; post-graduate students and researchers who need assistance in publishing empirical research in academic journals. Using their corpus of approximately 8 million words (The Specialized Corpus of Civil Engineering Research Articles), they conducted an extensive research not only on 4-word bundles, but also on 3-word, 5-word and even 6-word bundles. Along with analyzing the use of lexical bundles, they also studied individual keywords that are peculiarly employed in the field of civil engineering. Keywords were extracted by comparing the word use of their corpus to a bigger reference corpus (Corpus of Contemporary American English subcorpora; written fiction, magazine and newspaper texts, totaling 290.4 million words). The keywords that were sought by Gilmore & Miller (2018) to hold pedagogic value to novice engineers were then compared to several established wordlists (General Service List, Academic Word List). They concluded that keyword and cluster analysis are complementary in many ways – "keyword lists highlight the propositional content which typifies civil engineering texts, while word bundles 'frame' that content (p. 14)".

Similarly, in Korea, Nam (2017) studied lexical bundles specifically in the discipline of nuclear science and engineering and compared it to a general academic reference corpus (Academic Subcorpora of British National Corpus). His NSE (Nuclear Science and Engineering) corpus was made up of 222 texts, 1 million words. Lexical bundles were defined in his study as 4-word sequences that occur at least 20 times per million words in at least five texts. The structural basis for his analysis was Biber *et al.*'s (1999) classification, also consulting Cortes (2004) and Chen & Baker (2010) in order to "cross-check the examples of academic lexical bundles and the classification organization (p. 176)." An interesting finding in the NSE corpus was that verb phrase-based bundles were the most frequent structure while noun phrase-based bundles were the most frequent in the reference academic subcorpora of BNC. A disciplinary peculiarity of NSE that deviates from the general structural characteristics of lexical bundles in academia was thus identified in the results. Writers in engineering can perhaps benefit by being aware of such discipline-specific writing conventions and strengthen their academic writing by more convincingly appealing to the disciplinary audience.

Chapter 3. Data and Methodology

3.1. Data

In the following sections, significant issues in designing and compiling a corpus for a corpus-based study are described. These include how the data source used for this study was selected, overall summary of the corpus, tools for making and processing the data, and a detailed explanation of which parts of the data had to be manually excluded.

3.1.1. Corpus design and compilation

The corpus used in this study was collected from research articles in international academic journals. Deciding from which journals the data should be collected was an important criterion since disciplinary comparisons are to be made in this study. Selecting journals was based on two criteria. The first was a needs analysis surveying a small number of undergraduate and graduate students. They were asked if there is an academic journal that they commonly refer to in studying or researching. This criterion was adopted in order to ensure the pedagogic value of this study. Collecting data from a source that is remote to students' needs would likely to lead to results that are of less pedagogic value. The second was high journal ranking, and this was adopted to secure the generality of the language accepted in the respective disciplinary community. The journals thus selected according to the two criteria were *Applied Linguistics* (applied linguistics), *Journal of Management* (management), *Energy & Environmental Science* (engineering) and *Microbiology* (microbiology). Research articles in the respective journals were downloaded via online access.

The structural development of the research articles listed in these journals were checked for comparability, that is, whether it followed what Swales (1990) has described as *hour-glass macro structure* where research articles narrow a subject overview to a specific research question which is resolved later by an experiment, followed by broadening to relate to a wider field.

As mentioned in the introduction, researchers in the subfields of pedagogy and second language acquisition consider the nativity and proficiency an important criterion that must be matched across writers or participants. This study, however, focuses on the disciplinary variation and therefore takes a rather lenient stance in terms of proficiency; the articles used as data in this research met the strict publication regulations of the internationally renowned academic journals. This guaranteed the relevant and adequate proficiency among authors of various articles and ensured comparability required for analysis according to the direction and the aim of this thesis.

At least 500,000 words were collected per discipline for a total of 2.3 million words. Data was collected starting from the most recent issue to the one that exceeds 500,000 words. Table 4 below summarizes the information about the word count, text number, collection term, volume, and issue.

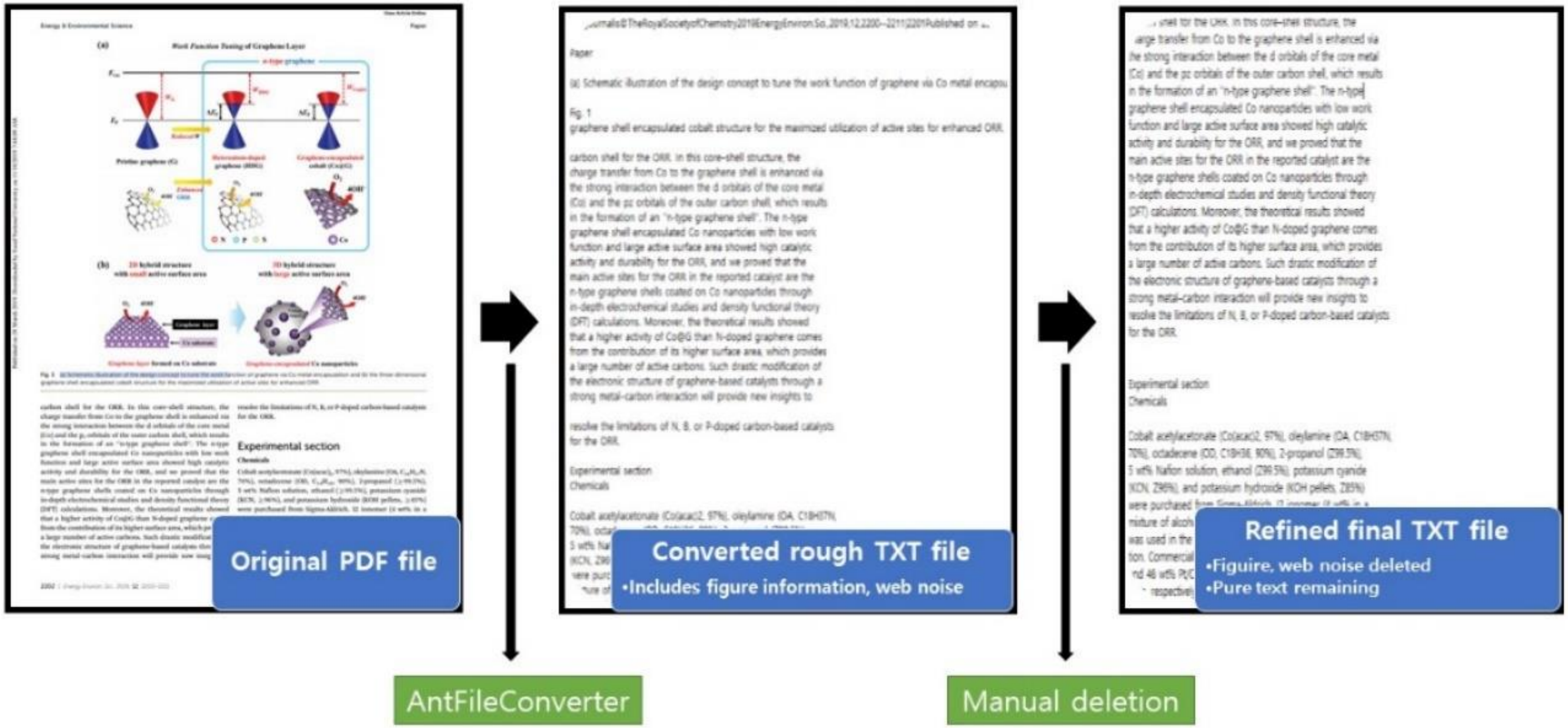
Table 4. Summary of the corpus

Discipline (Journal)	**Number of words**	**Number of articles**	**Collection term**	**Volume (Issue)**
Applied linguistics ***(Applied Linguistics)***	544,930	73	Dec 2017 – Aug 2019	38(6) – 40(4)
Management ***(Journal of Management)***	642,857	61	Mar 2019 – Jul 2019	45(3) – 45(6)
Engineering ***(Energy & Environmental Science)***	601,848	106	Jan 2019 – Jul 2019	12(1) – 21(7)
Microbiology ***(Microbiology)***	537,648	105	Sep 2018 – Aug 2019	164(9) – 165(8)
Total	2,327,283	345		

Again, the definite number of words and texts were matched as similarly as possible across disciplines, but there are slight discrepancies. It could further be inferred from the above table that the average text length, for example, is different across disciplines; management articles were longer in average than engineering texts because while the word count is relatively similar, significantly more engineering articles had to be collected in order to reach the same aim word count. Comparability is guaranteed by normalization and adopting the percentage dispersion criterion, rather than absolute number of texts in identifying lexical bundles.

3.1.2. Data refinement

Every research article beginning with the ones from the latest issue in the above-mentioned journals were downloaded in PDF form. The files were then converted into TXT form using Anthony's (2014) AntFileConverter. A manual refinement process was then necessary in order to exclude any part of the article that is not the author's language. Figure 1 below visualizes the refinement process. First, all web noise that occurred in the conversion process was deleted. These include copyright information, journal access information, article title and author name that appear at every page transition point. Next, all texts that gave additive information about the article or the author, such as author affiliation, contribution, funding, conflicts of interest, references and appendices were deleted. Visual information like figures and tables were also not linguistic information and were deleted. Finally, equations and formulas that were extremely common in engineering and



...urnals©TheRoyalSocietyofChemistry2019EnergyEnviron.Sci.,2019,12,2200–2211|2201Published on ...

Paper

(a) Schematic illustration of the design concept to tune the work function of graphene via Co metal encapsu

Fig. 1
graphene shell encapsulated cobalt structure for the maximized utilization of active sites for enhanced ORR.

carbon shell for the ORR. In this core–shell structure, the charge transfer from Co to the graphene shell is enhanced via the strong interaction between the d orbitals of the core metal (Co) and the pz orbitals of the outer carbon shell, which results in the formation of an "n-type graphene shell". The n-type graphene shell encapsulated Co nanoparticles with low work function and large active surface area showed high catalytic activity and durability for the ORR, and we proved that the main active sites for the ORR in the reported catalyst are the n-type graphene shells coated on Co nanoparticles through in-depth electrochemical studies and density functional theory (DFT) calculations. Moreover, the theoretical results showed that a higher activity of Co@G than N-doped graphene comes from the contribution of its higher surface area, which provides a large number of active carbons. Such drastic modification of the electronic structure of graphene-based catalysts through a strong metal–carbon interaction will provide new insights to

resolve the limitations of N, B, or P-doped carbon-based catalysts for the ORR.

Experimental section
Chemicals

Cobalt acetylacetonate (Co(acac)2, 97%), oleylamine (OA, C18H37N, 70%), octad...
5 wt% Na...
(KCN, ≥9...
were purc...
...ture of...

... shell for the ORR. In this core–shell structure, the ...arge transfer from Co to the graphene shell is enhanced via ...he strong interaction between the d orbitals of the core metal (Co) and the pz orbitals of the outer carbon shell, which results in the formation of an "n-type graphene shell". The n-type graphene shell encapsulated Co nanoparticles with low work function and large active surface area showed high catalytic activity and durability for the ORR, and we proved that the main active sites for the ORR in the reported catalyst are the n-type graphene shells coated on Co nanoparticles through in-depth electrochemical studies and density functional theory (DFT) calculations. Moreover, the theoretical results showed that a higher activity of Co@G than N-doped graphene comes from the contribution of its higher surface area, which provides a large number of active carbons. Such drastic modification of the electronic structure of graphene-based catalysts through a strong metal–carbon interaction will provide new insights to resolve the limitations of N, B, or P-doped carbon-based catalysts for the ORR.

Experimental section
Chemicals

Cobalt acetylacetonate (Co(acac)2, 97%), oleylamine (OA, C18H37N, 70%), octadecene (OD, C18H36, 90%), 2-propanol (≥99.5%), 5 wt% Nafion solution, ethanol (≥99.5%), potassium cyanide (KCN, ≥96%), and potassium hydroxide (KOH pellets, ≥85%) were purchased from Sigma-Aldrich. ...
mixture of alcoh...
was used in the...
tion. Commercial...
nd 46 wt% Pt/C...
respectively...

Figure 1. Data refinement process

microbiology had to be excluded for the same reason.

All visual information in the left-most original PDF file was converted textually in a sloppy fashion, so manual deletion was necessary, resulting in a pure text file as in the right-most final TXT file. For example, on the top of every page, copyright information, page number, author name and article title appeared. These were considered web noise and had to be deleted manually as well.

3.2. Methodology

In the following subsections, how the target lexical bundle list was made is first described in detail. In specific, extraction, identification, and exclusion of 4-word lexical bundles are addressed in order. Next, the ways in which the structural and functional classification of lexical bundles were adopted or modified is explained. The reasons for adopting or modifying previous classifications are also addressed in detail.

3.2.1. Lexical bundle extraction

Based on the refined data, a list of 4-word lexical bundles was extracted first. The scope of this research is limited to 4-word lexical bundles. There are several reasons for this, the most important being manageable size. Chen & Baker (2010) state that 4-word bundles are “found to be the most researched length for writing studies, probably because the number of 4-word bundles is often within a manageable size for manual categorization and concordance checks (p. 32).” Also, categorizing and displaying the structure and function of lexical bundles are clearer in 4-word bundles compared to 3-word bundles. Moreover, it is proven in Biber *et al*. (1999) that 3-word sequences are likely to lack significance owing to its commonality and that bundles over 5-words are somewhat rare to conduct research. Simpson-Vlach & Ellis (2010), however, argued that 3-word bundles were in fact discovered to comprise the majority of significant lexical bundles. A more comprehensive study including 3-word bundles could therefore lead us to richer results and implications, but as mentioned above, manageable size contained the scope to 4-word bundles.

Two tools were used for extraction; WordSmith Tools 7 and AntConc 3.5.7. Two complementary tools were used to double check for any discrepancies that may arise from differences in software processing or software malfunctioning. For example, certain mathematical signs used in engineering could not be processed in AntConc, resulting in fewer word counts than in WordSmith Tools. The *cluster* function was used in WordSmith Tools and *n-gram* function was used in AntConc to retrieve a list of 4-word lexical bundles.

3.2.2. Lexical bundle identification

The retrieved list, however, further needed to be distilled into a purer list of 4-word sequences that could technically be called lexical bundles. Excluding 4-word sequences that are irrelevant to this study and leaving only the relevant ones in interest was next in hand. For example, 4-word sequences occurring adjacently only a few times by chance could not be seen as lexical bundles according to this study. As briefly introduced in Chapter 2, such "how commonly used" criterion (cut-off frequency) and a "how widely used" criterion (dispersion) were adopted in the identification process. The cut-off frequency sets the minimum occurrences of bundles for them to be included in the analysis. Previous studies in the literature have employed various cut-off frequencies in target bundle identification, depending on whether the research attempts to comprehensively cover all identifiable bundles possible or qualitatively look into a relatively short and conservative list of what can undoubtedly be called lexical bundles. Biber *et al*. (1999) adopted the cut-off frequency of ten times per million words while Chen & Baker (2010) set a conservative cut-off frequency of 25 occurrences per million words and Cortes (2004) 20 times per million words. For 4-word bundles, a cut-off frequency of 40 times per million words is likely to be conservative or relatively high (Biber & Barbieri, 2007). Along this line, this research sets a conservative cut-off frequency of 40 times per million words. Since the word count of each subcorpus does not reach one million words, the cut-off frequency had to be normalized: 22 occurrences in applied linguistics and microbiology, 24 occurrences in engineering, and 26 occurrences in management. The second criterion, dispersion, is used to prevent idiosyncratic bundles (e.g. a bundle used repetitively only in one article) from appearing on the target bundle list. In the literature, the dispersion criterion is also employed in a somewhat arbitrary fashion; Biber & Barbieri (2007), Biber *et al*. (2004), Chen & Baker (2010), Cortes (2004) adopted the three to five text criterion while Hyland (2008b) used the percentage (10%) criterion. Since the number of texts in each discipline differ in this study, the percentage criterion (10%) was used to guarantee comparability among disciplines. This required bundles to appear in at least seven texts in applied linguistics, six texts in management, 11 texts in engineering and ten texts in microbiology. This was also a conservative criterion compared to the three to five text criterion since the least dispersion was six texts (management) and the most 11 (engineering).

In sum, both the cut-off frequency and dispersion criteria were as conservative as possible to produce lexical bundles that occur widely in relatively high frequency. This retrieved only the bundles that appeared at least 40 times per million words in at least 10% of the texts in each subcorpus as target bundles.

3.2.3. Exclusion criteria

The retrieved list still needed further manual exclusion because it included some irrelevant bundles. It turned out that 57 bundles were irrelevant to the purpose of this study, for the reasons summarized in Table 5. The full list of excluded bundles is provided in Appendix E. The exclusion criteria were established by amending Choi's (2015) and Salazar's (2011) criteria.

Table 5. Exclusion criteria

Exclusion Criteria	**Examples**
Topic-specific (context-dependent) bundles	*as a foreign language,* *x ray photoelectron spectroscopy,* *the electrochemical performance of,* *the wild type strain,* *methods bacterial strains and*
Bundles with acronyms	*density functional theory DFT,* *transmission electron microscopy TEM*
Bundles with proper nouns	*in the United States*
Bundles with numbers	*shown in Fig 1,* *at 37 c for,* *listed in Table 1*
Idioms	*state of the art*
Bundles with measuring units	*100 mg ml 1*
Fragments of bigger bundles	*used in this study,* *in this study are,* *this study are listed,* *study are listed in,* *are listed in table*
Bundles with possessive *'s*	*the manufacturer s instructions*

Web noise	*in the online supplement,* *available in the online*

Topic-specific (context-dependent) bundles were excluded because comparing a bundle that is exclusively used in one particular discipline (e.g. *the electrochemical performance of*) to other disciplines is less likely to yield meaningful results. Bundles with proper nouns were excluded for the same reason. Also, the only idiom that appeared in the corpus, *state of the art* was excluded since it cannot be considered a lexical bundle, according to the defining characteristics of lexical bundles mentioned in Section 2.1. above. That is, the idiosyncratic and perceptually salient meaning of the whole strain is newly created as a chunk, not derived by adding up the meanings of individual words. Furthermore, fragments of a 7-word bundle (*used in this study are listed in table*) appeared on the microbiology target list as 5 different 4-word bundles (*used in this study, in this study are, this study are listed, study are listed in, are listed in table*). This resulted in unnecessarily inflating the list of target bundles and was not included in the study since it was considered a 7-word bundle instead of five separate 4-word bundles. Strings that included non-word components (i.e. acronyms, numbers, measuring units, possessive *s*) were excluded as well. Some web noise (e.g. *available in the online*) was intermixed in the content, surviving the manual deletion procedure. Manual exclusion of these were necessary since they are not the author's language.

Overlapping bundles were also checked manually to avoid the inflation of the target list according to the type of overlap Chen & Baker (2010) referred to as *complete subsumption*. Two bundles are in the relation of complete subsumption if the occurrence of one bundle completely subsumes all occurrences of the other bundle. For example, the bundle *the end of the* occurred 27 times and the bundle *at the end of* occurred 15 times, all of which were included in the 27 occurrences of the former. After a concordance check, overlap of this kind is listed as one entity, with the additional entity of the less frequent bundle in parentheses (e.g. *(at) + the end of the*).

3.2.4. Structural classification

As introduced in Chapter 2, the structural classification of lexical bundles in Biber *et al.*'s (1999) research is adopted in this study. All categories were preserved without modification, apart from the *Pronoun/noun phrase + be (+...)* category being deleted due to its absence. Table 6 shows the 11 structures of lexical bundles which are further grouped based on part-of-speech.

Table 6. Structural classification of lexical bundles (Biber *et al.*, 1999 modified)

Grouping	**Structure**	**Example**
NP-based	Noun phrase with *of*-phrase fragment (NP + *of*)	*the nature of the,* *the effect of the*
	Noun phrase with other post-modifier fragments (NP)	*the fact that the,* *the degree to which*
PP-based	Prepositional phrase with embedded *of*-phrase fragment (PP + *of*)	*in the case of,* *in terms of the*
	Other prepositional phrase (fragment) (PP)	*on the other hand,* *at the same time*
VP-based	Anticipatory *it* + verb phrase/adjective phrase (Anticipatory *it*)	*it is possible that,* *it is important to*
	Passive verb + prepositional phrase fragment (Passive verb)	*is shown in fig,* *was used as the*
	Copula *be* + noun/adjective phrase (Copula *be*)	*is due to the,* *is consistent with the*
	(Verb phrase +) *that*-clause fragment ((V) + *that*)	*these results suggest that,* *that the relationship between*
	(Verb/adjective +) *to*-clause fragment ((V/A) + *to*)	*are more likely to,* *is negatively related to*
Adverbial clause fragment (Adv. cl.)		*as shown in fig,* *as can be seen*
Other expressions (Others)		*as well as the,* *play a role in*

The groupings based on part-of-speech was originally not introduced in Biber *et al.* (1999). Further structural studies of lexical bundles adopting their classification (e.g. Salazar 2010; Choi 2015; Nam 2017) introduced broader groupings of structure based on part-of-speech.

Noun phrase structures with and without “of” were grouped into *NP-based* grouping and the same grouping applied to the two prepositional phrase structures with and without "of". There were five structural correlates that could be grouped into the *VP-based* grouping and adverbial clause fragments were classified independently.

3.2.5. Functional classification

While structural studies of lexical bundles in the literature uniformly adopt Biber *et al.*'s (1999) classification, the functional taxonomy diverges into adopting either Biber *et al.*'s (2004), or Hyland's (2008b) taxonomy, depending mainly on the genre characteristics of the corpus used. As discussed in Chapter 2, Biber *et al.* (2004) collected corpus data not only from academic writing, but also from genres like textbooks, conversation, and classroom teaching. Their functional taxonomy, therefore, incorporates the bundles identified in spoken data and non-academic writing as well, while Hyland's (2008b) research used data solely composed of academic writing. It was therefore decided that Hyland's (2008b) functional taxonomy is more appropriate for the aim of this study.

By referring to Salazar (2010), several modifications were necessary for the aim of the present thesis. First, the *topic* function was deleted since topic-specific bundles were excluded. Also, the *transition*, *resultative signals* functions were separated since transition signals are described as "establishing *additive* or *contrastive* links between elements" and resultative signals "marking *inferential* or *causative* relations between elements". The coordinated terms "additive and contrastive" and "inferential and causative" seemed reasonably distinct so they were separated. Several new functions (grouping, citation, generalization, objective) introduced in Salazar (2010) were deleted since it seemed that their functional description could sufficiently be incorporated in the description of other functions. The *acknowledgement* function was deleted since they were absent; manual deletion of the data eliminated all acknowledgement, contribution, funding sections of research articles. Table 7 summarizes the resulting functional taxonomy to be adopted for this study.

Table 7. Functional taxonomy modified from Hyland (2008b)

Function	Description (Examples)
Research-oriented	Helps writers to structure their activities and experiences of the real world
Location	Indicate time/place (*at the beginning of, at the same time*)
Procedure	Indicate events, actions and methods (*the use of the, the role of the*)
Quantification	Indicate measures, quantities, proportions and changes thereof (*a wide range of, the magnitude of the*)
Description	Indicate quality, degree and existence (*the structure of the, the size of the*)
Text-oriented	Concerned with the organization of the text and its meaning as a message or argument
Additive	Establish additive links between elements (*in addition to the*)

Contrastive	Compare and contrast different elements (*on the other hand, as compared with*)
Inferential	Mark inferential relations between elements (*these results suggest that*)
Causative	Mark causative relations between elements (*as a result of, the results of the*)
Structuring	Text-reflexive markers which organize stretches of discourse or direct reader elsewhere in the text (*in the present study*)
Framing	Situate arguments by specifying and limiting conditions (*in the case of, with respect to the*)
Participant-oriented	These are focused on the writer or reader of the text
Stance	Convey the writer's attitudes and evaluations (*are likely to be, may be due to*)
Engagement	Address readers directly (*it should be noted that, as can be seen*)

As can be seen in the table above, the three broad functional orientations (research, text, participant) were preserved. Text-oriented functions were separated more specifically to further capture the subtle functional roles that lexical bundles play in academic discourse.

As Biber *et al*. (2004) mentions, a single bundle may serve more than one function in various contexts. They state that bundles like *at the end of*, for example, can serve multifunctionally as a time reference, place reference or text deictic reference. Hyland (2008b) also adopts this concept of multifunctionality and classifies bundles like *in the present study*, for example, functioning both as research-oriented location bundle and text-oriented structuring signal. The present study also adopts this concept of multifunctionality.

Chapter 4. Analysis and Results

This chapter is an arrangement of the similarities and differences of the use of lexical bundles that were found in the corpus of the current study across the four disciplines in terms of frequency, structure and function. First of all, the frequency, represented by types and tokens of target bundles, will be analyzed.

4.1. Comparison of frequency

Adopting the conservative cut-off frequency and dispersion criteria resulted in retrieving a relatively small number of target bundles compared to the results of other recent corpus-based studies. The retrieved bundles, however, are indisputable lexical bundles that are used frequently across various authors in the discipline. Table 8 below summarizes the type, token and normalized per million-word frequency of lexical bundles in each subcorpus.

Table 8. Types and tokens of 4-word lexical bundles in each subcorpus

Discipline	Types (per million-words)	Tokens (per million-words)	Total words
Applied linguistics	26 (48)	894 (1,640)	544,930
Management	41 (64)	1,814 (2,822)	642,857
Engineering	42 (70)	2,048 (3,403)	601,848
Microbiology	29 (54)	1,316 (2,448)	537,648
Total	138	6,072	2,327,283

Types indicate the number of distinct bundles, and tokens indicate the number of total bundles. Since corpus size is slightly different across disciplines, it is necessary to compare the normalized per million-word frequency in the parentheses.

Engineering noticeably outnumbered all other disciplines in both types and tokens. Token count was more than double the token count of applied linguistics and type was also the highest. This result is in accordance with Hyland's (2008b) findings in which the electrical engineering subcorpora in his study used bundles most frequently than other disciplines. He stated that in engineering, there is "greater reliance on pre-fabricated structures (p. 9)" which he speculated to be "a consequence of the relatively abstract and graphical nature of technical communication (p. 9)." In fact, looking into the engineering research articles in this study revealed the tendency of engineers to frequently use figures and develop the text by regularly referring to them. This tendency is supported by the first ranked bundle in engineering *as shown in fig*, which overwhelmingly stands out in the list of target bundles (548 occurrences per million words) as shown in Table 9. Bundles shown in all four disciplines are boldened and three disciplines italicized.

Table 9. Top 20 lexical bundles in each discipline (per million words)

Applied linguistics		**Management**		**Engineering**		**Microbiology**	
Bundle	**Freq.**	**Bundle**	**Freq.**	**Bundle**	**Freq.**	**Bundle**	**Freq.**
on the other hand	166	the extent to which + (the)	297	as shown in fig	548	in the presence of + (the/a)	498
at the same time	116	are more likely to	235	as a function of + (the)	213	in the absence of	244
the ways in which	81	in the context of	107	are shown in fig	201	in this study we	134
on the basis of	73	*on the basis of*	107	is shown in fig	133	**on the other hand**	113
in the case of	72	**as well as the**	106	**in the case of**	123	the expression of the	102
in the context of	72	*as a result of*	100	**on the other hand** + (the)	120	**in the case of**	100
it is important to	72	it is important to	86	**as well as the**	106	as shown in fig	86
in the present study	70	be more likely to	72	(higher/lower) + than that of the	93	**as well as the**	84
the meaning of the	68	it is possible that	70	(which) + is consistent with the	88	was added to the	78
as well as the	66	are less likely	67	in the	75	was used as a	73

		to		presence of			
in terms of the	61	the nature of the	65	was used as the	75	has been shown to + (be)	67
in the form of	61	**in the case of**	64	to the formation of + (the/a)	73	it is possible that	63
of the variance in	59	are likely to be	62	with respect to the	71	by the addition of	60
in relation to the	53	at the individual level	61	in the range of	68	was found to be	60
the extent to which	53	in terms of the	59	in this work we	66	as compared to the	54
the nature of the	53	**on the other hand**	59	to that of the	65	in the regulation of	54
the results of the	51	the indirect effect of	59	(can) + be attributed to the	61	in the present study	50
(at) + the end of the	50	as shown in table	58	can be used to	61	these results suggest that	50
as a result of	46	on the relationship between + (the)	58	in good agreement with + (the)	61	to be involved in	48
in this article we	46	the validity of the	58	as shown in the	58	the fact that the	46

It was also interesting to see that similar to the exceptional *as shown in fig* in engineering, an exceptional lexical bundle(s) was identified in all other fields but applied linguistics. Specifically, the top two bundles in management were used 297 and 235 times respectively, but the use of the third bundle remained at 107. In engineering, the 548 occurrences of the top bundle are followed by the 213 uses of the second bundle. The gap between the first and second bundle is also wide in microbiology (498 vs. 244). In applied linguistics, however, the difference between bundles is relatively small, with the first bundle being used 166 times and the second 116. This absence of an exceptional bundle in applied linguistics is one reason why applied linguistics shows the least number of tokens. The reason for this is hard to say but a speculative possibility based on checking the characteristics of the data is that linguistic research articles dealt with a wider range of applied subfields with different focuses: translation, education, assessment, acquisition, and so on, while other disciplines had a relatively concentrated goal that most research articles converge on. The writing conventions thus seemed more solidified towards a more common interest in management, engineering, and microbiology, resulting in the emergence of some prototypical lexical bundles.

There were only three bundles occurring in all four disciplines (the boldened *on the other hand, in the case of, as well as the*) and four bundles appearing in three disciplines (the italicized *on the basis of, in the form of, as a result of, the fact that the*). This lack of shared bundles and discipline-specific distribution of lexical bundles was also found in Hyland's (2008b) study, in which only five bundles were shared across all four disciplines and 14 bundles shared in three disciplines. Moreover, over half of the lexical bundles in each discipline of Hyland's (2008) study did not occur in any other discipline, with 30% of the bundles in each discipline appearing in two other fields. Similarly, in this study, a great number of bundles were not shared and used uniquely. The ratio of discipline-unique bundles was 36% in applied linguistics, 70% in management, 69% in engineering and 62% in microbiology. The significantly low ratio in applied linguistics can again speculatively be addressed by the reasons mentioned above; other disciplines had a more concentrated goal across articles, while applied linguistics described a relatively wide range of linguistic phenomena, displaying a more general picture in lexical bundle use. Nevertheless, 36% is still a high ratio because it means over one-third of all target bundles were unshared. Hyland (2008b) identified more shared bundles than this study, but considering he adopted a relatively lenient cut-off frequency of 20 occurrences per million words, it could roughly be argued that both studies support the discipline-specificity in the display of lexical bundles.

It was also interesting to find that the three bundles occurring in all disciplines in this study (*in the case of, on the other hand, as well as the*) were also found to be shared in all disciplines in Hyland's (2008b) research. Two of these three bundles (*in the case of* and *on the other hand*) were found in Biber *et al.*'s (1999) research to be the most frequently occurring 4-word bundles in academic prose, appearing over 100 times per million words.

Looking into the similarities in cognate fields also known in the literature as the dichotomous distinction, soft sciences (applied linguistics, management) and hard sciences (engineering, microbiology) provides insight to some noteworthy tendencies and will also be dealt in detail in the following sections.

4.2. Comparison of structure

Table 10 summarizes the structural classification of lexical bundles in each discipline following Biber *et al.*'s (1999) taxonomy. Overall, PP-based bundles were used most frequently and the second most frequently used structure differed across disciplines. The results and analysis will be addressed respectively in terms of the part-of-speech groupings: NP(noun phrase)-based, PP(prepositional phrase)-based, VP(verb phrase)-based. The numbers in Table 10 indicate types (distinct occurrences). Bundles that were ambiguous to classify into a single structural classification were checked in detail, using the concordance function of both software (WordSmith Tools, AntConc). The context in which the lexical bundle occurred was manually checked to determine which structural correlate the bundle better fits into.

Table 10. Structural classification of lexical bundles by discipline following Biber *et al*.'s (1999) taxonomy

	Structure	Applied ling.	Management	Engineering	Microbiology
NP-based	Noun phrase + of	4 (15.4%)	7 (17.1%)	5 (11.9%)	3 (10.3%)
	Other noun phrases	3 (11.5%)	4 (9.8%)	2 (4.8%)	2 (6.9%)
	Total	7 (26.9%)	11 (26.9%)	7 (16.7%)	5 (17.2%)
PP-based	Prep. phrase + of	8 (30.8%)	8 (19.5%)	12 (28.6%)	7 (24%)
	Other prep. phrases	7 (26.9%)	7 (17.1%)	7 (16.7%)	3 (10.3%)
	Total	15 (57.7%)	15 (36.6%)	19 (45.3%)	10 (34.3%)
VP-based	Anticipatory it + verb/adj.	1 (3.8%)	2 (4.9%)	2 (4.8%)	1 (3.4%)
	Passive verb + prep. phrase fragment	1 (3.8%)	0 (0)	7 (16.7%)	4 (13.8%)
	Be + noun/adj. phrase	0 (0)	0 (0)	2 (4.8%)	1 (3.4%)
	That-clause fragment	0 (0)	1 (2.4%)	0 (0)	1 (3.4%)
	To-clause fragment	1 (3.8%)	10 (24.3%)	1 (2.4%)	3 (10.3%)
	Total	3 (11.5%)	13 (31.6%)	13 (28.7%)	10 (34.3%)
	Adverbial phrases	0 (0)	1 (2.4%)	3 (7.1%)	2 (6.9%)
	Others	1 (3.8%)	1 (2.4%)	1 (2.4%)	2 (6.9%)
	Total	26 (100%)	41 (100%)	42 (100%)	29 (100%)

4.2.1. NP-based bundles

NP-based bundles were more frequently used in the soft sciences than in the hard science disciplines. A comprehensive list of all the NP-based bundles is given in Table 11.

Table 11. NP-based bundles in each subcorpus

	Noun phrase + *of*	**Other noun phrases**
Applied linguistics	*the meaning of the* *the nature of the* *the results of the* *(at) + the end of the*	*the ways in which* *the extent to which* *the fact that the*
Management	*the nature of the* *the indirect effect of* *the validity of the* *the moderating effect of* *the positive effect of* *our understanding of the* *the negative effects of*	*the extent to which + (the)* *the degree to which* *the negative relationship between* *the positive relationship between*
Engineering	*the effect of the* *the end of the* *the inset of fig* *a wide range of* *the performance of the*	*an increase in the* *the fact that the*
Microbiology	*the expression of the* *a large number of* *the effect of the*	*the fact that the* *an increase in the*

It can be seen from the use of NP-based bundles that management articles are heavily concerned with studying the relationship between variables as shown in the NP bundles used only in management like *the negative relationship between, the positive relationship between*. Whether one variable has an effect on other variables seem to be an issue of interest, as shown in the use of NP-based bundles including multiple kinds of "effect" (*the indirect / moderating / positive / negative effect of*). Below are some examples of how they were used specifically in context.

(1) We find partial support for the moderation effects of property rights protection on *the positive relationship between* humane orientation and CBA performance. (MAN45(4)_1)

(2) To verify the significance of *the indirect effect of* the interaction of the HPW system and organizational change on creativity via collective learning, we performed bootstrapping with Mplus, number of bootstrap samples (n=1,000; Hayes, 2015) (MAN45(3)_3)

The 4-word bundle *the effect of the* was also identified in engineering and microbiology, but with significantly less tokens. Furthermore, management studies exclusively focused more specifically on the various types of effects and bundles including *relationship* was absent in other disciplines, as shown in Table 11.

4.2.2. PP-based bundles

PP-based bundles were most frequently used in all four disciplines. Table 12 lists all PP-based bundles in each discipline.

Table 12. PP-based bundles in each subcorpus

	Prep. phrase + *of*	**Other prep. phrases**
Applied linguistics	*on the basis of* *in the case of* *in the context of* *in terms of the* *in the form of* *as a result of* *over the course of + (the)* *at the beginning of*	*on the other hand* *at the same time* *in the present study* *of the variance in* *in relation to the* *in this article we* *in the current study*
Management	*in the context of* *on the basis of* *as a result of* *in the case of* *in terms of the* *in the form of* *on the role of* *in the face of*	*at the individual level* *on the other hand* *on the relationship between + (the)* *at the same time* *to the extent that* *of the relationship between + (the)*
Engineering	*as a function of + (the)* *in the case of* *on the other hand + (the)* *(higher/lower) + than that of the* *in the presence of* *to the formation of + (the/a)* *in the range of* *to that of the* *as a result of*	*with respect to the* *in this work we* *in good agreement with + (the)* *in addition to the* *in contrast to the* *as a result the* *for the first time*

	in the form of *on the basis of* *on top of the*	
Microbiology	*in the presence of* + *(the/a)* *in the absence of* *in the case of* *by the addition of* *in the regulation of* *to that of the* *with the exception of*	*in this study we* *on the other hand* *in the present study*

Overall, PP-based bundles including "of" was more frequent than the ones without it. Although the above table shows that engineering used PP-based bundles the most, it is necessary to focus on the rate and proportion of PP-based bundles used. Applied linguistics stands out in that PP-based bundles were used at an extremely high rate (58%). Many PP-based bundles used in applied linguistics were related to logical relations between propositions, preventing leaps and protecting coherence as in (3).

(3) Combined with this growth, and partly *as a result of* it, the number of references in each paper has increased consistently through the 20th century. (LIN40(1)_4)

The lexical bundle in interest is in between two propositions, linking them and protecting logical coherence which turns out to be an important aspect in this particular linguistic research article.

4.2.3. VP-based bundles

Overall, linguists significantly used less VP-based bundles than authors in other disciplines. All VP-based bundles are listed in Table 13. The axles are converted for visual comfort.

Table 13. VP-based bundles in each subcorpus

	Applied ling.	**Management**	**Engineering**	**Microbiology**
Anticipatory *it*	*it is important to*	*it is important to* *it is possible that*	*it should be noted + (that)* *it can be seen + (that)*	*it is possible that*
Passive verb	*it can be seen*		*are shown in fig* *is shown in fig* *was used as the* *(can) + be attributed to the* *is attributed to the* *can be found in* *are presented*	*was added to the* *was used as a* *was found to be* *was observed in the*

			in fig	
Copula *be*			*(which) + is consistent with the* *is due to the*	*be due to the*
***That*-clause fragment**		*that the relationship between + (the)*		*these results suggest that*
***To*-clause fragment**	*can be used to*	*are more likely to* *be more likely to* *are less likely to* *are likely to be* *is more likely to* *is positively related to* *is negatively related to* *have been found to* *likely to engage in* *we were able to*	*can be used to*	*has been shown to + (be)* *to be involved in* *have been shown to*

The most significant distinction in the table above and in all structural comparison was that bundles of passive structure were dominant in the hard sciences and almost absent in soft sciences, as shown again in Figure 2.

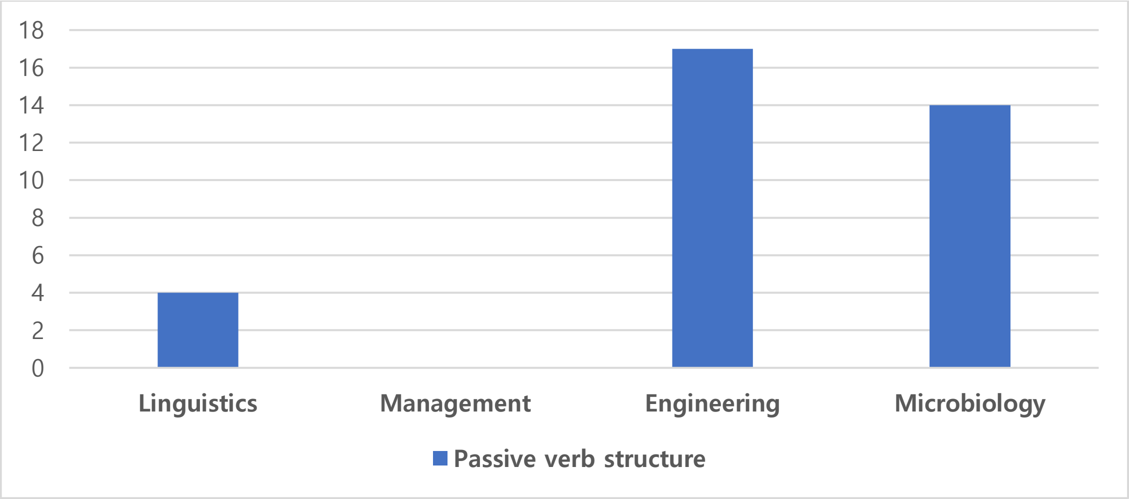


Figure 2. Passive structure in each discipline (per cent)

There were no instances in management in which lexical bundles of passive structure were used and only one instance in applied linguistics, yielding a proportion of a mere 4%. On the other hand, engineering stood out with 17% of all retrieved bundles having passive structure. Microbiology also showed a high proportion (14%) of passive structures.

Looking into the concordance, this seemed to reflect engineers' and biologists' preference to hide authorial voice in describing the experimental methodology (4)-(5) and reporting results (6).

(4) The electrocatalyst was coated onto glassy carbon (GC) as the working electrode, lithium foil *was used as the* counter electrode, Pt foil *was used as the* reference electrode, and 10 mM Li2S6 solution *was used as the* electrolyte. (ENG12(1)_16)

(5) One-sixth volume of bromophenol blue and glycerol *was added to the* protein samples before they were subjected to electrophoresis at constant current (MIB165(4)_1)

(6) It should be noted that at the later stages of growth (>35 h), some recovery of growth *was observed in the* case of a few clones, which could have been due to the accumulation of revertants in the M. (MIB165(7)_1)

In (4) and (5), both authors are describing their experimental methodology, but no agent role is shown on the surface. The author in (6) is similarly hiding his or her authorial voice in describing the findings of the experiment.

"The relatively abstract and graphical nature of technical communication" mentioned above in the frequency comparison section also played a role in the dominance of passive structure in engineering and microbiology, as bundles directing to visual information were structured as passives, as in (7).

(7) The comparisons of the rate capability and cycling stability *are shown in Fig*. 7b. (ENG12(7)_8)

When authors attempt to draw the audience's attention elsewhere to certain visual information such as figures in (7), as well as tables, equations, appendices and so on, this role was frequently conveyed by using lexical bundles of passive structure.

Another interesting result is that the use of bundles including a *to*-clause fragment structure in management is unrivaled (24%, while other disciplines remain at maximum 10%). Specifically, bundles indicating likeliness were exclusively used in management, as in (8)-(9) (e.g. *are more likely to, be more likely to, are less likely to, are likely to be*).

(8) Those practices can provide the rich, nurturing environment where employees *are more likely to* initiate risk-taking behaviors and, in turn, boost collective learning (MAN45(3)_3)

(9) Group members *are less likely to* exert effort to contribute to the group when they feel free of social monitoring. (MAN45(3)_11)

It is worth mentioning here that structure and function of lexical bundles are closely related to each other. The above *to*-clause fragment structure indicating tendency almost always functions as impersonally expressing the writers' stance.

This is a participant-oriented function which is predominantly used in management and the reason will be discussed in the functional comparison section below.

Another factor contributing to the high occurrence of *to*-clause structures in management is that the main interest of social science research is to disclose the positive or negative relationships between social elements. This is in alignment with the use of NP-based bundles indicating the relationship between and effect of variants (e.g. *the negative relationship between, the positive effect of*). Consequently, in management papers, there were various occurrences of *to*-clause bundles in hypotheses sections, as in (10)-(12).

(10) Hypothesis 2b: A higher performance proven goal orientation in top executives *is positively related to* firm environmental scanning. (MAN45(5)_7)

(11) Hypothesis 1: State ownership *is negatively related to* firm financial performance. (MAN45(6)_2)

(12) Hypothesis 3: CEO-board chair separation *is negatively related to* corporate misconduct. (MAN45(6)_10)

Almost every management paper included such hypothesis sections with lexical bundles referring to relationships or effects between social variables.

4.3. Comparison of function

The final feature to be analyzed is the function of the bundles, that is, the role they play in academic discourse. Table 14 summarizes the functional classification by discipline based on a modified version of Hyland's (2008b) taxonomy.

4.3.1. Research-oriented bundles

Applied linguistics used research-oriented bundles at the highest rate (30%). Second is engineering (26%) in which research-oriented bundles were quite frequently used compared to the low rate in management (18%) and microbiology (19%). As for specific disciplinary variation, bundles serving multifunctionally both as description and quantification abounded in the soft sciences, as in (13)-(14).

> (13) An important assessment of the value of a practical theory is *the extent to which* it can ask new and different questions on both the practice under investigation and other existing theories about the practice. (LIN39(1)_2)
>
> (14) To measure collectivism in the home country, we used organizational in-group collectivism, which captures "*the degree to which* individuals express pride, loyalty, and cohesiveness in their

organization" (House *et al*., 2004: 46) (MAN45(4)_1)

Only text-oriented and participant-oriented multifunctional lexical bundles were found in the hard sciences and such research-oriented multifunctional lexical bundles could only be found in applied linguistics and management.

Table 14. Functional classification of lexical bundles by discipline based on a modified version of Hyland’s (2008b) taxonomy

Function		**Applied ling.**	**Management**	**Engineering**	**Microbiology**
Research-oriented	Location	2 (6.7%)	0 (0)	3 (6.5%)	0 (0)
	Procedure	1 (3.3%)	1 (2%)	2 (4.3%)	2 (6.3%)
	Description	5 (16.7%)	5 (9.8%)	4 (8.7%)	3 (9.4%)
	Quantification	1 (3.3%)	3 (5.9%)	3 (6.5%)	1 (3.1%)
	Total	9 (30%)	9 (17.7%)	12 (26%)	6 (18.8%)
Text-oriented	Additive	2 (6.7%)	2 (3.9%)	3 (6.5%)	3 (9.4%)
	Contrastive	1 (3.3%)	7 (13.7%)	6 (13%)	3 (9.4%)
	Inferential	1 (3.3%)	7 (13.7%)	1 (2.2%)	3 (9.4%)
	Causative	2 (6.7%)	7 (13.7%)	7 (15.2%)	6 (18.8%)
	Structuring	4 (13.3%)	1 (2%)	7 (15.2%)	4 (12.5%)
	Framing	7 (23.3%)	7 (13.7%)	5 (10.9%)	5 (15.6%)
	Total	17 (56.6%)	31 (60.7%)	29 (63%)	24 (75.1%)
Participant-oriented	Stance	3 (10%)	10 (19.6%)	2 (4.3%)	2 (6.3%)
	Engagement	1 (3.3%)	1 (2%)	3 (6.5%)	0 (0)
	Total	4 (13.3%)	11 (21.6%)	5 (10.8%)	2 (6.3%)
Total		30 (100%)	51 (100%)	46 (100%)	32 (100%)

It is also worth highlighting that applied linguistics used description bundles at the highest rate (17%, while other disciplines remain under 10%). Concordance view yielded interesting results regarding the description bundle, *the meaning of the*, which is exclusively used in applied linguistics. Out of the 37 occurrences of this bundle, 11 were those referring literally to the word "word", giving 11 occurrences of the 5-word bundle *the meaning of the word*. Other expressions were also used to point to the meaning of a particular word in the text, such as *noun, verb,* and *risks*. Including these, it resulted in a total of 17 occurrences of this bundle pinpointing one word and further elaborating on the meaning of it as in (15).

> (15) The presence of risk and uncertainty (uncertainties) of Pattern 8 enriches *the meaning of the* 'risks', and hence enables *the meaning of the* whole pattern to be conveyed in a more straightforward way. (LIN39(3)_1)

It could be inferred that the sophisticated, detailed meaning or nuance of individual words tend to play a crucial role in linguistic articles, so authors were likely to elaborate on potentially ambiguous word meanings in detail using lexical bundles with such functions.

In terms of research-oriented bundles, Hyland's (2008b) research found results that differ from this research. The hard knowledge fields showed a greater concentration of research-oriented bundles (49% in electrical engineering, 48% in biology, 31% in applied linguistics and 36% in business studies). In this study, however, there was almost no distinction in the use of research-oriented bundles between the soft and hard sciences (average 24% in soft sciences and 23% in hard sciences). This distinction perhaps rises from the different characteristics of the corpus makeup, that is, Hyland's (2008b) corpus included master theses and PhD dissertations as well as research articles in his data, while the corpus collected for this study solely comprise research articles. Some characteristic features of research articles, contrasted to student writing, may have influenced the results. This will be dealt in detail in the next text-oriented function section.

4.3.2. Text-oriented bundles

Perhaps the most exceptional tendency in the functional analysis of bundles is the prevalence of text-oriented bundles overall. This is closely related to the characteristics of research articles as discussed by Hyland (2008a) in his study comparing the structural and functional patterns of dissertations written by students and research articles written by professionals in the field. He argued that research articles written for publication preponderate with text-oriented bundles because they need to interact with a literature, bring warrants, build background and connect ideas. They also guide the audience around the text and point out limitations. They intend to lead readers to the author's analysis of research process and results, signaling the main outcomes to be derived from the research and highlight the inferences from which the author wants the audience to draw in the argument. This is different from student writing which is generally aimed at displaying their knowledge to a strictly restricted party or an individual who would likely to be grading them. The content mainly being discussed in student writing tend to be already-known knowledge which the student attempts to prove and convince that he or she is familiar with. Bringing in warrants and citing renown, established studies and thus interacting with the literature is less likely to happen in student writing, compared to professional academic writing. Sisilia *et al*.'s (2019) study on student writing supports this, finding that research-oriented lexical bundles were used the most in student writing. Specifically, students most commonly adopted bundles that indicate procedure and description of their research. The results in this study shown above in Table 14 is in agreement with Hyland's (2008a) results in that text-oriented bundles are by far the most frequently used type of bundles. Figure 3 displays this in graph presentation.

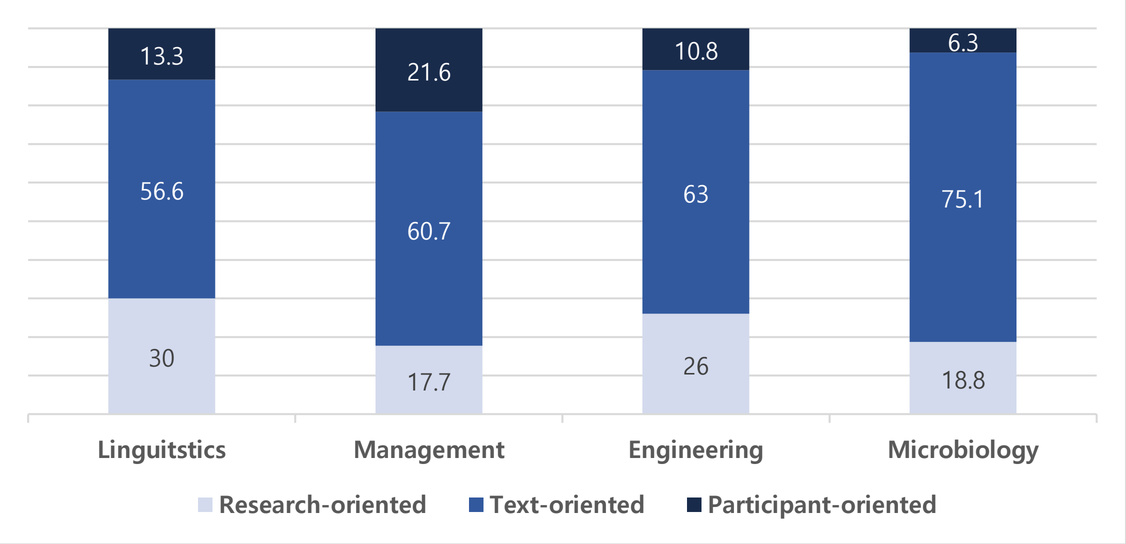


Figure 3. Functional distribution of lexical bundles by discipline (per cent)

Noticeably, in microbiology, over 75% of all identified lexical bundles were used as text-oriented functions. The other three disciplines also display this dominance of text-oriented bundles. Combining the proportion of research and participant-oriented bundles cannot reach the overwhelming use of text-oriented bundles in all disciplines.

Causative and inferential functions occurring the most in microbiology is in accordance with Hyland's (2008b) research, in which resultative (inferential plus causative) bundles were also found most frequent in his biology subcorpora.

In all discipline, framing and structuring bundles were used quite frequently. Research articles in academic journals attempt to convince the readers to approve and concur with the arguments made by the author. Articles are written and published not by recurrently displaying well-known knowledge, but rather by building on previous knowledge and ultimately proposing original ideas to the professional readership. Structuring and framing bundles observed in this study play a crucial role in accomplishing goals of this kind. Structuring bundles direct the readers to different points of the text, drawing their attention to the desired points at the desired order, as in (16)-(17). Framing bundles, as stated in the definition, situate arguments by specifying and limiting conditions and thus pave the road that leads to the inferences and conclusions the author is willing the readers would make along with him or her, as in (18).

(16) *It can be seen* from the above equations that the colour of the solar cell, i.e. its spectral reflectance, affects both the photocurrent density Jph (eqn (2) and (4)) and the recombination current density Jrec0 (eqn (5)-(7)) roughly proportionally. (ENG12(4)_3)

(17) *As shown in Fig*. S5, the PI substrate that is used has a porous surface microstructure that can promote the discharge of H2 produced in the ECD reaction and can thereby effectively prevent corrugation and stripping of the copper coating. (ENG12(1)_3)

(18) In the current study, it was clear that LD measures which were computed on texts that were lemmatized *on the basis of* the word family were less powerful in predicting CEFR levels as well as different Pearson scores than those that were computed *on the basis of* the lemma. (LIN39(3)_2)

As the audience reads (16), their attention is likely to be drawn to the equation that the author is referring to. Likewise, reading (17) naturally leads the reader's attention to the particular figure that the author wants the readers to focus on at the current stage of the author's argument development. At the beginning of (18), a relatively broad proposition which could potentially be interpreted ambiguously or misleadingly is presented. The author then uses a framing lexical bundle *on the basis of* to specify a limiting condition, disbranching the undesirable ambiguous or misleading interpretations and leaving only the intended interpretations that would desirably lead to the author's main implications in the end.

4.3.3. Participant-oriented bundles

Participant-oriented bundles are the means by which authors interact with readers in academic discourse. A dialogic relationship can be established through these bundles. Noticeably, management used participant-oriented bundles, especially stance, at a rate exceptionally higher than any other discipline (21.6%). This was closely related to the *to*-clause fragment structure indicating likeliness. While engineers and biologists were likely to prefer hiding their authorial voice in describing experiment methodology and reporting results, professionals in management studies seemed not to bother in displaying their stance. This supports Hyland's (2008b) results in which researchers in social sciences overwhelmingly employed bundles indicating their stance. He explains the reason for this may be because in social sciences, "writers have to establish their claims through more explicit evaluation and engagement: personal credibility, and explicitly getting behind arguments, plays a far greater part in convincing discourse for these writers (p. 18)." It was thus relatively frequent for writers in management to convey their attitudes and evaluations in forms such as *are more likely to, it is important to, it is possible that*.

It should be noted that although authors in management showed their stance more often than writers in other disciplines, this is done impersonally and indirectly. Bundles including personal pronouns such as *I* or *we* do appear on the target bundle list (e.g. *in this study / work we, in doing so we, we were able to*) but with significantly less types and tokens than impersonal stance bundles. Such depersonalization strategies are known to show the author's attempt to weaken personal bias or subjectivity when expressing their ideas. The indirectness in the expression of authors' stance can be explained to "largely convey a reluctance to express complete commitment to a proposition, allowing writers to present information as an opinion rather than accredited fact (p. 18)" (Hyland, 2008b). These indirect expressions, also known as hedges in the literature of pragmatics, protect the author's argument from being face-threatened by diminishing the author's agent role. Below are examples in management in which writers are employing the depersonalization (19) and indirectness (20) strategy in presenting arguments.

> (19) Highly committed founders and loyal employees *are more likely to* endure the difficulties that start-ups face, thus improving the prospects of organizational survival. (MAN45(3)_6)
>
> (20) Given that System 2 may offer some advantages over default processing, *it is important to* understand how to facilitate the activation of System 2 to mitigate the potential irrational, harmful, or biased outcomes of System 1. (MAN45(6)_3)

While displaying the writer's stance, the agent role is not shown in both the above examples. Also, the strength of which the author's argument is being delivered is attenuated in (20). It leaves a hatch of academic responsibility for the author since it is not in a form of definite argument.

4.4. Summary of findings

First in terms of frequency, types and tokens of lexical bundles were both the highest in engineering. The following disciplines were management, microbiology, and applied linguistics in order. It was speculated that engineers rely relatively more on prefabricated linguistic structures due to the abstract and graphical nature of technical communication. In fact, figures played a crucial role in engineering research articles as shown by authors developing their text by frequently using and referring to figures. Also, a discipline-unique distribution of lexical bundles was found in all four disciplines. The ratio of such bundles used exclusively within discipline was surprisingly high in engineering (70%) and relatively low in applied linguistics (36%). This supports Hyland's (2008b) results in which approximately 50% of the bundles identified in his research were found to be unshared across disciplines. It was also interesting that the three lexical bundles identified in all four disciplines of this study (*in the case of, on the other hand, as well as the*) were also found to be shared in all four disciplines of Hyland's (2008b) research. Furthermore, Biber *et al.* (1999) found that two of these (*in the case of, on the other hand*) were the most frequently used lexical bundle in academic discourse. Structurally, PP-based lexical bundles were used the most in all disciplines. Passive structure was used significantly more in disciplines of the hard knowledge fields (engineering, microbiology) than soft knowledge fields (applied linguistics, management). Passive structure abounded in describing the experimental methodology and reporting results in engineering and microbiology articles. Also, management papers noticeably focused on unearthing the relationship between social variables. That is, whether one variable has an impact on another was mainly at the heart of management articles. For example, four different lexical bundles including *effect* (*the indirect / moderating / positive / negative effect of*) and two different bundles including *relationship* (*the negative / positive relationship between*) was exclusively identified in management. It was also interesting that PP-based bundles were used at the highest rate in applied linguistics (58%). Most PP-based bundles used in applied linguistics were related to logical relations, preventing leaps and protecting coherence. Functionally, text-oriented bundles were used by far the most in all four disciplines. It was addressed that characteristics of research articles which comprise the data of this study influenced such results. Also, authors in management used participant-oriented bundles, especially stance bundles, significantly more than other disciplines. This was because authors in management tried more explicitly to deliver their argument as opposed to authors in engineering and microbiology. Although showing stance, writers in management commonly employed the depersonalization and indirectness strategies in order to weaken personal bias or subjectivity when expressing their ideas. In engineering and microbiology papers, however, authors were likely to hide authorial stance.

Chapter 5. Implications and Conclusion

This study aimed to discover how similarly, or differently lexical bundles are used in academic research articles across four different academic disciplines (applied linguistics, management, engineering, microbiology). Three main features (frequency, structure and function) were studied and few similarities were discovered while many disciplinary peculiarities were found. The most noteworthy overarching tendency observed was that functionally, text-oriented lexical bundles were used predominantly in all disciplines. The reason for this was speculated to be due to the nature of research articles. Research articles are different from student writing in which research-oriented lexical bundles are used more frequently than text-oriented bundles. Another similarity was that prepositional phrase based lexical bundles were the most common structure in all four disciplines. In terms of frequency, only three lexical bundles were found to occur in all four disciplines and four in three disciplines.

Contrary to such few similarities, many disciplinary peculiarities and differences were found. Lexical bundles not occurring in other disciplines and used uniquely only within a single discipline abounded, with management studies showing a vast 70% proportion of discipline-unique bundles. Subsequent proportions by discipline were 69% (engineering), 62% (microbiology), and 36% (applied linguistics). Furthermore, the most distinctive peculiarity was perhaps the passive structure being actively used in the hard science disciplines and almost absent in the soft science disciplines. Authorial stance was closely related to this tendency since authors in engineering and microbiology preferred hiding it, describing experiments and reporting results in passive structure. Writers in management, on the other hand, were less reluctant to hiding authorial stance in order to strengthen their argument. Also, PP-based bundles were overwhelmingly used in applied linguistics and this was attributed to linguists being more careful in connecting logical relations.

As mentioned in Chapter 2, this study is based on Hyland's (2008b) framework, adopting its functional taxonomy and the disciplines. Results that are in agreement with Hyland's (2008b) study were found, but some contrary results were also found. First, in terms of agreeing results, engineering displayed the widest range of distinct and total lexical bundles in both studies. The difference between engineering and other disciplines in this research was especially significant. Also, both studies proved a unique disciplinary distribution of lexical bundles, with only a few bundles appearing across disciplines. In Hyland's (2008b) study, over 50% of all target lexical bundles in each discipline did not appear in other disciplines. The numbers were more dramatic in the present study; management and engineering almost reached 70% unique distribution and applied linguistics a relatively less 36%. Moreover, passive structure was preferred in engineering and microbiology in both studies. This preference for passive structure in the hard knowledge fields was one of the most significant differences found in this study. Finally, both studies confirmed that lexical bundles showing author's stance was predominantly used in social sciences.

Hyland's (2008b) research, however, revealed that there is a functional difference of lexical bundles between the soft and hard knowledge fields: research-oriented bundles used more in the hard knowledge fields and text-oriented bundles used more in the soft knowledge fields. This is different from the results of this study since text-oriented lexical bundles were overall the most common in all disciplines and rather used more frequently in the hard sciences. The difference in corpus makeup, namely the text type being homogeneous in the current study and heterogeneous in Hyland's (2008b) is considered to have an impact on such contrary results. Also, PP-based lexical bundles were used significantly more in the social science discipline of Hyland's (2008b) data, but the difference was insignificant across disciplines in this study. Moreover, Hyland's (2008b) study revealed that the *anticipatory it* structure was used more frequently in the hard sciences, but this study found that it was used more in the soft science disciplines instead.

The results of this study could be somewhat disappointing to general English instructors since not many similarities encompassing all disciplines were found. Rather, a discipline-specific distribution of lexical bundles was observed with structural and functional characteristics also being reasonably distinct in each discipline or in the soft/hard knowledge field distinction. The discipline-specific characteristics of this kind, disciplinary specificity, or disciplinary literacy can also be valuable especially to a focused target audience, because as Hyland (2009) states, "academic writing is only effective when writers use conventions that other members of their community find familiar and convincing (p. 5).” Being aware of such discipline-specific strategies of using lexical bundles, rather than the general strategies, can assist learners and apprentice writers of a discipline enhance their writing by successfully appealing to the target audience and agreeing to the conventions that the disciplinary community find convincing, especially in the EAP (English for Academic Purposes) and ESP (English for Specific Purposes) context.

As for limitations, the data used for this research may not adequately represent the disciplinary fields that are studied because the sample is not big enough and is collected from a narrow source (one journal per discipline). Also, the arbitrary criteria in lexical bundle identification could result in inferences to lack comparability. Finally, widening the scope of analysis to include 3-word, and 5-word bundles could lead us to more significant cross-disciplinary findings.

References

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*(2), 81-92.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis and Applications* (pp.101-122). Oxford: Oxford University Press.

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Biber, D. (2006). *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: Benjamin.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286.

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

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

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

Choi, B. (2015). Lexical bundles in linguistics research articles: A contrastive study of native and non-native writing (Unpublished master's thesis). Seoul National University, Seoul.

Cortes, V. (2002b). Lexical bundles in academic writing in history and biology. (Unpublished doctoral dissertation). Northern Arizona University, United States.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397-434.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education, 17,* 391-406.

Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing, 16*(3), 129-147.

Durrant, P. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics, 38*(2), 165-193.

Erman, B., & Warren, B. (2000). The idiom principle and the open-choice principle. *Text, 20*, 29–62.

Firth, J. R. (1951). Modes of meaning. *Papers in Linguistics, 1934-1951* (pp.118-149). London: Oxford University Press.

Gilmore, A., & Millar, N. (2018). The language of civil engineering research articles: A corpus-based approach. *English for Specific Purposes, 51*, 1-17.

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

Granger, S., & Meunier, F. (Eds.). (2008a). *Phraseology: An Interdisciplinary Perspective* Amsterdam: John Benjamins.

Haswell, R. (1991). *Gaining Ground in College Writing: Tales of Development and Interpretation.* Dallas: Southern Methodist University Press.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language.* London: Routledge.

Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41-62.

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21.

Hyland, K. (2009). Writing in the disciplines: Research evidence for specificity. *Taiwan International ESP Journal, 1*(1), 5-22.

Jalali, Z. S., & Moini, M. R. (2014). Structure of lexical bundles in introduction section of medical research articles. *Procedia-Social and Behavioral Sciences, 98*, 719-726.

Kim, J. (2018). The effects of learning lexical bundles on the college students in an English composition class. *Korean Journal of Applied Linguistics, 34*(1), 3-27.

Kwary, D., A., & Ratri, D., & Artha, A., F. (2017). Lexical bundles in journal articles across academic disciplines. *Indonesian Journal of Applied Linguitics, 7*(1), 131-140.

McCulley, G. (1985). Writing quality, coherence, and cohesion. *Research in the Teaching of English, 19*(3), 269-282.

Nam, D. (2017). Lexical bundle structures of nuclear science and engineering research article. *Language Facts and Perspectives, 40*, 167-186.

Nattinger, J., & DeCarrico, J. (1992). *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Noguchi, J. (2006). *The Science Review Article: An Opportune Genre in the Construction of Science*. Bern: Peter Lang.

O’Keeffe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching.* Cambridge: Cambridge University Press.

Salazar, D. (2011). Lexical bundles in scientific English: A corpus-based study of native and non-native writing. (Unpublished doctoral dissertation). Universitat de Barcelona, Spain.

Scott, M., & Tribble, C. (2006). *Textual Patterns: Keywords and Corpus Analysis in Language Education.* Amsterdam: John Benjamins.

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

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4), 487-512.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sisilia, A. D. I., & Sri, W. F., & Januarius, M. (2019) Structure and function of lexical bundles in the literature review of undergraduate studnets' final projects. *English Education Journal, 9*(1), 62-73.

Siyanova-Chanturia, A., & Martinez, R. (2015) The idiom principle revisited. *Applied Linguistics, 36*(5), 549-569.

Stubbs, N., & Barth, I. (2003). Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Language, 10*(1), 61-104.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Wang, Y. (2017). Lexical bundles in spoken academic ELF genre and disciplinary variation. *International Journal of Corpus Linguistics, 22*(2), 187-211.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics, 21*(4), 463-489.

Wray, A. (2002). *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

Wright, H. R. (2019). Lexical bundles in stand-alone literature reviews: Sections, frequencies, and functions. *English for Specific Purposes, 54*, 1-14.

Appendices

A. Complete list of target bundles in applied linguistics

Rank	Lexical bundle	Freq.	Dispersion	Structure	Function1	Function2
1	on the other hand	91	55%(40)	PP	Contrastive	
2	at the same time	63	45%(33)	PP	Additive	
3	the ways in which	44	26%(19)	NP	Description	
4	on the basis of	40	25%(18)	PP + of	Framing	
5	in the case of	39	33%(24)	PP + of	Framing	
6	in the context of	39	32%(23)	PP + of	Framing	
7	it is important to	39	38%(28)	Anticipatory it	Stance	Engagement
8	in the present study	38	21%(15)	PP	Structuring	
9	the meaning of the	37	25%(18)	NP + of	Description	
10	as well as the	36	34.%(25)	Others	Additive	
11	in terms of the	33	26%(19)	PP + of	Framing	
12	in the form of	33	25%(18)	PP + of	Framing	
13	of the variance in	32	12%(9)	PP	Framing	
14	in relation to the	29	25%(18)	PP	Framing	
15	the extent to which	29	23%(17)	NP	Description	Quantification
16	the nature of the	29	25%(18)	NP + of	Description	
17	the results of the	28	18%(13)	NP + of	Causative	
18	(at) + the end of the	27	22%(16)	NP + of	Location	
19	as a result of	25	22%(16)	PP + of	Causative	
20	in this article we	25	21%(15)	PP	Stance	Structuring
21	over the course of + (the)	25	16%(12)	PP + of	Procedure	
22	the fact that the	24	25%(18)	NP	Description	
23	can be used to	23	18%(13)	(V/A) + to	Inferential	Stance

24	at the beginning of	22	19%(14)	PP + of	Location	
25	can be seen in	22	14%(10)	Passive	Structuring	
26	in the current study	22	14%(10)	PP	Structuring	

B. Complete list of target bundles in management

Rank	Lexical bundle	Freq.	Dispersion	Structure	Function1	Function2
1	the extent to which + (the)	191	79%(48)	NP	Description	Quantification
2	are more likely to	151	69%(42)	(V/A) + to	Stance	Inferential
3	in the context of	69	57%(35)	PP + of	Framing	
4	on the basis of	69	43%(26)	PP + of	Framing	
5	as well as the	68	57%(35)	Others	Additive	
6	as a result of	64	44%(27)	PP + of	Causative	
7	it is important to	55	44%(27)	Anticipatory it	Stance	Engagement
8	be more likely to	46	34%(21)	(V/A) + to	Stance	Inferential
9	it is possible that	45	39%(24)	Anticipatory it	Stance	Inferential
10	are less likely to	43	36%(22)	(V/A) + to	Stance	Inferential
11	the nature of the	42	36%(22)	NP + of	Description	
12	in the case of	41	34%(21)	PP + of	Framing	
13	are likely to be	40	41%(25)	(V/A) + to	Stance	Inferential
14	at the individual level	39	33%(20)	PP	Framing	
15	in terms of the	38	38%(23)	PP + of	Framing	
16	on the other hand	38	36%(22)	PP	Contrastive	
17	the indirect effect of	38	20%(12)	NP + of	Causative	
18	as shown in table	37	28%(17)	Adv. cl.	Structuring	
19	on the relationship between + (the)	37	28%(17)	PP	Contrastive	
20	the validity of the	37	10%(6)	NP + of	Description	
21	in the form of	36	36%(22)	PP + of	Framing	

22	is more likely to	36	34%(21)	(V/A) + to	Stance	Inferential
23	the moderating effect of	35	20%(12)	NP + of	Causative	
24	the degree to which	34	31%(19)	NP	Quantification	Description
25	the positive effect of	34	20%(12)	NP + of	Causative	
26	the negative relationship between	33	20%(12)	NP	Contrastive	
27	on the role of	31	30%(18)	PP + of	Procedure	
28	is positively related to	30	23%(14)	(V/A) + to	Contrastive	
29	our understanding of the	30	30%(18)	NP + of	Stance	
30	the negative effects of	30	18%(11)	NP + of	Causative	
31	at the same time	29	31%(19)	PP	Additive	
32	that the relationship between + (the)	29	30%(18)	(V) + that	Contrastive	
33	in the face of	28	18%(11)	PP + of	Framing	
34	is negatively related to	28	23%(14)	(V/A) + to	Causative	
35	to the extent that	27	26%(16)	PP	Description	Quantification
36	have been found to	26	25%(15)	(V/A) + to	Causative	
37	in doing so we	26	33%(20)	PP	Stance	
38	likely to engage in	26	25%(15)	(V/A) + to	Inferential	
39	of the relationship between + (the)	26	30%(18)	PP	Contrastive	
40	the positive relationship between	26	18%(11)	NP	Contrastive	
41	we were able to	26	26%(16)	(V/A) + to	Stance	

C. Complete list of target bundles in engineering

Rank	Lexical bundle	Freq.	Dispersion	Structure	Function1	Function2
1	as shown in fig	330	74%(78)	Adv. cl.	Structuring	
2	as a function of + (the)	128	43%(46)	PP + of	Causative	
3	are shown in fig	121	57%(60)	Passive	Structuring	
4	is shown in fig	80	49%(52)	Passive	Structuring	
5	in the case of	74	32%(34)	PP + of	Framing	
6	on the other hand + (the)	72	38%(40)	PP + of	Contrastive	
7	as well as the	64	38%(40)	Others	Additive	
8	(higher/lower) + than that of the	56	31%(33)	PP + of	Contrastive	
9	(which) + is consistent with the	53	35%(37)	Copula be	Contrastive	
10	in the presence of	45	24%(25)	PP + of	Framing	
11	was used as the	45	22%(23)	Passive	Procedure	
12	to the formation of + (the/a)	44	26%(28)	PP + of	Description	
13	with respect to the	43	30%(31)	PP	Framing	
14	in the range of	41	24%(25)	PP + of	Quantification	
15	in this work we	40	25%(27)	PP	Stance	
16	to that of the	39	25%(26)	PP + of	Contrastive	
17	(can) + be attributed to the	37	23%(24)	Passive	Causative	
18	can be used to	37	24%(25)	(V/A) + to	Inferential	Stance
19	in good agreement with + (the)	37	255(27)	PP	Contrastive	
20	as shown in the	35	22%(23)	Adv. cl.	Structuring	
21	it should be noted + (that)	35	23%(24)	Anticipatory it	Engagement	
22	is attributed to the	33	22%(23)	Passive	Causative	
23	can be found in	31	22%(23)	Passive	Structuring	
24	an increase in the	30	15%(16)	NP	Procedure	Description
25	as a result of	30	16%(17)	PP + of	Causative	
26	in the form of	30	20%(21)	PP + of	Framing	

27	it can be seen + (that)	30	15%(16)	Anticipatory it	Structuring	Engagement
28	on the basis of	29	19%(20)	PP + of	Framing	
29	the effect of the	29	21%(22)	NP + of	Causative	
30	the end of the	29	14%(15)	NP + of	Location	
31	the inset of fig	29	16%(17)	NP + of	Location	
32	are presented in fig	28	15%(16)	Passive	Structuring	
33	in addition to the	28	20%(21)	PP	Additive	
34	in contrast to the	28	20%(21)	PP	Contrastive	
35	is due to the	28	23%(24)	Copula be	Causative	
36	on top of the	28	14%(15)	PP + of	Additive	Location
37	a wide range of	26	18%(19)	NP + of	Quantification	
38	the fact that the	26	16%(17)	NP	Description	
39	the performance of the	26	14%(15)	NP + of	Description	
40	as a result the	25	20%(21)	PP	Causative	
41	for the first time	25	12%(13)	PP	Quantification	
42	as can be seen	24	13%(14)	Adv. cl.	Engagement	

D. Complete list of target bundles in microbiology

Rank	Lexical bundle	Freq.	Dispersion	Structure	Funtion1	Function2
1	in the presence of + (the/a)	268	62%(65)	PP+of	Framing	
2	in the absence of	131	43%(45)	PP+of	Framing	
3	in this study we	72	44%(46)	PP	Stance	Structuring
4	on the other hand	61	29%(30)	PP	Contrastive	
5	the expression of the	55	22%(23)	NP+of	Description	
6	in the case of	54	25%(26)	PP+of	Framing	
7	as shown in fig	46	25%(26)	Adv. cl.	Structuring	
8	as well as the	45	35%(37)	Others	Additive	
9	was added to the	42	28%(29)	Passive	Additive	
10	was used as a	39	23%(24)	Passive	Procedure	
11	has been shown to + (be)	36	24%(25)	(V/A) + to	Causative	

12	it is possible that	34	36%(27)	Anticipatory it	Stance	Inferential
13	by the addition of	32	20%(21)	PP + of	Additive	
14	was found to be	32	22%(23)	Passive	Causative	
15	as compared to the	29	10%(11)	Adv. cl.	Contrastive	
16	in the regulation of	29	14%(15)	PP + of	Framing	
17	in the present study	27	19%(20)	PP	Structuring	
18	these results suggest that	27	17%(18)	(V) + that	Inferential	
19	to be involved in	26	17%(18)	(V/A) + to	Inferential	
20	the fact that the	25	22%(23)	NP	Description	
21	was observed in the	25	18%(19)	Passive	Structuring	
22	a large number of	24	17%(18)	NP + of	Quantification	
23	play a role in	24	16%(17)	Others	Causative	
24	the effect of the	23	12%(13)	NP + of	Causative	
25	an increase in the	22	13%(14)	NP	Procedure	Description
26	be due to the	22	16%(17)	Copula be	Causative	
27	have been shown to	22	18%(19)	(V/A) + to	Causative	
28	to that of the	22	18%(19)	PP + of	Contrastive	
29	with the exception of	22	16%(17)	PP + of	Framing	

*Structure abbreviations

NP + *of* → Noun phrase with *of*-phrase fragment

NP → Noun phrase with other post-modifier fragment

PP + *of* → Prepositional phrase with embedded *of*-phrase fragment

PP → Other prepositional fragment

Anticipatory *it* → Anticipatory *it* + verb phrase / adjective phrase

Passive → Passive verb + prepositional phrase fragment

Copula *be* → Copula *be* + noun phrase / adjective phrase

(V) + *that* → (Verb phrase +) *that*-clause fragment

(V/A) + *to*	→	(Verb / adjective +) *to*-clause fragment
Adv. cl.	→	Adverbial clause fragment
Others	→	Other expressions

E. Full list of excluded bundles

	Applied linguistics	**Management**	**Engineering**	**Microbiology**
Topic-specific(context-dependent) bundles	*as a foreign language*		*X ray photoelectron spectroscopy, energy dispersive X ray, the electrochemical performance of, the electronic structure of, on the surface of, the thickness of the, the density of the, at a current density, an order of magnitude*	*to the wild type + (strain), of the wild type, in the wild type, a final concentration of, compared to the wild, according to the manufacturer, the wild type and, strains were grown in, the plates were incubated, cells were grown in, methods bacterial strains and, under the control of, to a final concentration*
Bundles with acronyms			*ray photoelectron spectroscopy XPS, X ray diffraction XRD, density functional theory DFT, transmission electron microscopy TEM, scanning electron*	

			microscopy SEM	
Bundles with proper nouns		*in the United States*		
Bundles with numbers			*shown in Fig 1, and Table s1 esi, in Table s1 esi, shown in Fig 4*	*f u ml 1, at 37 c for, an od 600 of, Table s1 available in + (the), listed in Table 1, 50 mg ml 1 + (and), were incubated at 37 + (c), to an od 600, h at 37 c + (in), 10 mg ml 1, and incubated at 37, listed in Table s1, at 37 c with*
Idioms			*state of the art*	
Bundles with measuring units				*c f u ml, 100 mg ml 1 + (and)*
Fragments of bigger bundles				*used in this study, in this study are, this study are listed, study are listed in, are listed in Table*
Bundles with possessive *‘s*				*the manufacturer s instructions*
Web noise		*in the online*		*available in the online*

		supplement		*+ version + of + this + article*

국문초록

4개 학문분야에서의 어휘다발에 대한 코퍼스 기반 연구

최봉준

서울대학교 대학원

영어영문학과 영어학 전공

본 논문에서는 4개의 서로 다른 학문 분야에서의 어휘다발 사용을 연구한다. 해당 학문 분야는 응용언어학, 경영학, 공학 그리고 미생물학인데 이 네 분야는 각각 인문사회, 사회과학, 자연과학, 생명과학의 특징을 적절히 반영한다고 판단하여 선택되었다. 분야 당 저명 국제 학술지에 실린 연구 논문들에서 최소 500,000 단어를 모아 총 약 250만 단어의 코퍼스를 구축하였다. 세 가지의 특징이 주로 연구되는데, 첫번째는 어휘다발의 빈도 비교이다. 다음으로 어휘다발의 사용이 학문분야간 어떻게 다르거나 유사한지 Biber *et al.* (1999)의 구조적 분류법에 따라 연구한다. 마지막으로는 Hyland (2008b)가 정의한 어휘다발의 기능 분류법에 따라 어휘다발의 기능을 분석한다. 빈도 비교의 결과, 공학분야가 다른

모든 분야보다 어휘다발(서로 다른 어휘다발, 총 어휘다발 모두)을 가장 많이 사용하는 것으로 밝혀졌다. 그리고 어휘다발은 대개 고유하게 한 학문분야 안에서만 사용되었다. 예를 들어, 공학에서 확인된 어휘다발 중 70%는 다른 세 분야에서 나타나지 않았다. 구조 비교의 결과, 자연과학, 생명과학 분야(공학, 미생물학)에서 인문사회 분야(응용언어학, 경영학)보다 수동태 구조를 눈에 띄게 많이 사용하는 것으로 드러났다. 저자의 입장이나 자세를 글에 드러내는 것이 이 경향성과 관련이 있었는데, 공학과 미생물학 논문의 저자들은 대개 겉으로 자신들의 입장을 드러내기를 꺼려한 반면, 경영학 논문의 저자들은 그렇지 않았다. 기능 비교의 결과, 텍스트 지향 기능을 하는 어휘다발이 네 학문분야 모두에서 가장 많이 사용되었다. 이러한 경향성은 학생들의 글쓰기와 대비되는 연구논문의 몇 가지 특징 때문인 것으로 예상되었다. 이와 같이 어휘다발을 여러 학문 분야에서 어떻게 다르게 (또는 유사하게) 사용하는지 연구함으로써 각 학문 공동체가 설득력 있고 익숙하다고 여기는 형식적인 어휘 관습을 발견할 수 있고, 이는 곧 그 학문분야의 특수성을 발견하는 것으로, 교실 환경에서 교육적으로 기여할 수 있을 것으로 예상된다.

주요어 : 어휘다발, 코퍼스 기반 연구, 구조, 기능, 학문 분야별 차이

학번 : 2018-28156