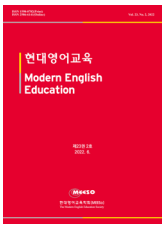
Modern English Education

https://doi.org/10.18095/meeso.2022.23.2.10



Semantic Properties of Lexical Bundles Used by Young Adult Korean EFL Students at Different Grade Levels*

Yu Kyoung Shin
Hallym University

ARTICLE INFO

Received: 30 April 2022
Revised: 29 May 2022
Accepted: 09 June 2022

Examples in: English
Applicable Languages: English
Applicable Levels: Secondary/Tertiary

KEYWORDS

Korean EFL student/
semantic prosody/
lexical bundle
한국인 영어학습자/
의미적 운율/
말뭉치

ABSTRACT

Shin, Yu Kyoung. (2022). Semantic properties of lexical bundles used by young adult Korean EFL students at different grade levels. *Modern English Education, 23*(2), 10-21.

This study explores the potential of lexical bundles (LBs) as a way for investigating academic writing development of EFL students. This study used a corpus of essays by Korean EFL students in high school or the first week of college divided by school-level group into three subcorpora (first-year high school; second-year high school; and third-year high school/incoming college). It investigated how groups used LBs in context by examining these LBs' semantic prosodies and preferences. Findings showed marked differences in structures and functions of bundles favored by first-year high school students versus two higher school-level groups, with the former using more bundles characteristic of colloquial spoken language. On the other hand, a contextual analysis of LBs shared across subcorpora showed more similarities between the first-year group and the second-year group regarding semantic prosodies and preferences. These two adjacent groups tended to project positive, neutral, and negative affective meanings via LBs at similar rates that differed from rates of the third-year/incoming college group. These results shed light on learners' developmental trajectory toward being proficient academic writers in terms of their use of formulaic multiword sequences in academic prose.

I. INTRODUCTION

A particular type of frequently occurring multiword sequence is called a lexical bundle (LB). LBs refer to groups of three or more words that recur in a given genre (Biber et al., 1999). Most of the existing studies on LBs are limited to academic writing produced by university students and professional academics; we know little about the extent to which student writers, in particular EFL learners, use academic discourse conventions *before* they enter university settings. This situation is partially due to the lack of systematically compiled nonnative corpora of secondary school student writing. The lack of available corpus data for this population is a serious gap in corpus-based studies. This comparison of LB uses by high school students with different English proficiencies will be useful in that the findings will help us understand developmental sequences of academic skills beginning earlier in students' academic careers; that is, starting in high school rather than in university.

One of the reasons it is important to compare LBs used by different groups of Korean students is that the population will exhibit particular and unique semantic prosodies

*This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A8060863).

Yu Kyoung Shin (Assistant Professor)
School of Global Studies, Hallym University, 1 Hallymdaehak-gil, Chuncheon-si, Kangwon-do, 24252 South Korea
Email: yshin@hallym.ac.kr

and preferences via LBs, based, for example, on their knowledge and the texts they have already encountered in educational and other settings. Most research on LBs in EFL student writing compares the students' production and usage of certain sets of bundles (based on the LBs known to be used by professional academic writers) with first language usage (Pérez-Llantada, 2014). The particular interest of this study is whether and to what extent Korean students studying English as a foreign language have access to and the ability to use English word sequences that frequently co-occur, and how Korean students understand and can utilize such sequences' semantic properties.

II. LITERATURE REVIEW

1. Lexical Bundles

Lexical bundles are sequences of three or more words that frequently occur in a given register; for example, *the beginning of the, it is important to,* and *due to the fact that* are common LBs in academic registers (Biber et al., 1999). Research has identified specific sets of bundles widely used in academic genres (e.g., Y. Kwon, 2021; E. Park, 2019; Salazar, 2014; Y. Shin et al., 2018, 2021), and demonstrated that the ability to use them competently is a key element of academic skills (e.g., Bychkovska & Lee, 2017; Y. Lee et al., 2020; Y. Shin & Y. Kim, 2017; Simpson-Vlach & Ellis, 2010).

Several researchers have exploited this specificity of LB repertoires to conduct comparative analyses of academic writing by different writer populations (e.g., Bychkovska & Lee, 2017; Paquot, 2017; Salazar, 2014). While this line of research has illuminated first language (L1) and second language (L2) usages of LBs in terms of types and textual functions, it has also raised new questions. Methodologically, most of the previous studies compare different kinds of academic writing, for instance, comparing published research articles and student essays (e.g., Wei & Lei, 2011). Yet distinct types of writing have different functions and must meet different expectations, which logically would influence deployment of LBs and other formulaic language (Pérez-Llantada, 2014; Pan et al., 2016; Y. Shin, 2019; Y, Shin, 2018).

Furthermore, there is a lack of research on EFL secondary school student writing (e.g., Northbrook & Conklin, 2019a, 2019b). Most of the prior research on this topic is concerned with higher education, and many existing corpora are composed of academic writing by university students (e.g., Y. Kwon, 2021; C. Yoon & J. Choi, 2015) and professional academics (e.g., Salazar, 2014; D. Shin & Y. Shin, 2020). The absence of systematically compiled corpora of EFL writers 'before' the university level has left a serious gap. This study's corpus collects a new body of data that informs our understanding of the early stages of the developmental sequences of Korean students' English academic writing skills, which will help us build a more complete picture of the trajectories students follow as they become proficient writers. In addition, the study's focus on this population's use of the formulaic language characteristic of argumentative essays provides detailed information on an aspect of language use that is closely intertwined with the development of proficiency.

2. Semantic Prosodies and Preferences

Another important reason for investigating English LB usage among Korean high school students is that we know very little about how this population's background knowledge (e.g., their knowledge of genres from their reading in Korean and English at this point in their education) affects their English production (e.g., Hyland, 2008). Because the bulk of LB research on L2 usage seeks to identify deviations from native speaker norms (Pérez-Llantada, 2014), most studies have neglected to ask interesting questions about variation in LB usage in context among L2 groups.

The study employs the tools of corpus linguistics to understand the interaction of LB usage and the content or meaning of the learners' texts. In particular, the study analyzes the LBs identified in the corpus in terms of semantic preference and semantic prosody. Semantic preference describes collocational preferences of individual items (e.g., words, phrases, LBs) within semantic categories or within lexical sets (e.g., Partington, 2004; Y. Shin, 2020; Sinclair, 2004; Stubbs, 2001).

Semantic prosody is also about collocational preferences but focuses on evaluative meaning. This part of the analysis is based on the observation that as language learners develop competence in a target language, they gain knowledge of native speakers' preferences for the co-occurrence of words or phrases (i.e., habitual collocations; Stubbs, 2001).

The role of semantic prosody has been explored in the past two decades by researchers including Oster (2010), Hunston (2011), and Cortes and Hardy (2013), Y. Shin (2020), for example, examined the semantic prosody of lexical bundles in academic essays written by native and nonnative English-speaking university students, and observed marked differences between the two language groups. The nonnatives were more likely than the natives to use lexical bundles in positively valenced contexts. For example, both groups used the LB *it would be the* in their direct responses to writing prompts. For the native writers, the examination of the noun phrases following this LB showed generally negative prosody, at a rate of 72.4%, while, for the nonnatives, the same LB was only seldomly used with negative prosody (12.5%). This trend, in fact, was observed more generally; that is, the native writers frequently made negative arguments (e.g., *disagree with the statement, do not agree with*) while the nonnative writers more often preferred positive arguments (e.g., *agree with the statement, so I agree with*).

Formulaic expressions like LBs are themselves a kind of collocation, but further collocate with the words and phrases that form their context. L2 mastery of formulaic language thus is intertwined with cumulative target language experience over time, just as vocabulary acquisition is (Partington, 2004). Because the acquisition of items

occurs through repeated encounters, the semantic associations and affective contexts in which items most frequently occur may be acquired along with the item. Yet the development of these associations is likely to take place differently for native and nonnative speakers, whose language histories and sociocultural knowledge differ (e.g., Moon, 1992; Morley & Partington, 2009; Y. Shin, 2020). For this reason, the present study is interested in whether and how young adult Korean EFL students' use of LBs is affected by habitual co-occurrences and semantic associations.

This study therefore addresses gaps in this research area by using a corpus of essays by Korean EFL students in high school and the first week of college, strictly matched for genre (i.e., argument essays) and writing prompt (i.e., identical topics and time constraints). The study's first step is to identify the lexical bundles used by the group. It next investigates the semantic prosodies and preferences of the bundles.

3. Present Study

This study examines how EFL students of different year-levels use LBs in terms of semantic prosodies and preferences in context. To do so, the study uses a corpus of essays produced by Korean EFL student writers divided into three subcorpora by the students' school-year level: first-year high school; second-year high school; and third-year high school/incoming college. (High school in Korea lasts three years; thus, the college student participants had recently completed their third and last year of high school.) These three subcorpora are completely comparable, as they are closely matched for writing prompts and topics. To address the two specific research questions below, the study first identifies lexical bundles in each subcorpus, and then analyzes their structures and functions (RQ1). It then examines the contextual usages of the shared bundles, that is, those found across subcorpora (RQ2).

1) In a corpus of English argumentative essays produced by Korean EFL students of different school years, what LBs occur most frequently?

2) How does each school-year group of students use the shared bundles in terms of semantic preference and semantic prosody?

III. METHOD

1. Corpus Data

The study uses a corpus of English argumentative essays written by Korean high school students and incoming college students (in the first week of their first semester of college). The study was originally designed to gather data from high schools only, with the researcher visiting participating schools to administer the essay tests in person.

However, due to COVID-19, data collection could not be completed as planned, with the greatest shortfall in the third-year high school students' essays (n = 24). This shortfall was partly made up by including essays on the same topics written by incoming college students, which had already been collected as part of an ongoing project. Thus, the third group combines those in their third (and final) year of high school and those in their first week of college. As this study aims to explore the developmental stages of Korean EFL students' English writing on their trajectory toward academic writing proficiency, the inclusion of the incoming college population was deemed reasonable.

TABLE 1
Description of Subcorpora

Corpus	Number of essays	Mean length	Total size
First-year	564	98.2	55,380
Second-year	211	124.6	26,285
Third-year+	140	330.7	46,293
Total	915	139.8	127,958

Note. The third-year+ corpus comprises essays by third-year high school students (n = 24, total words: 5,120) and incoming college students (n = 116, total words: 41,173).

As shown in Table 1, the corpus contains essays by over 900 students (139.8 words on average). They were instructed to write an argumentative essay on a given writing topic for 50 minutes; one of three writing prompts was given to each student. An example topic involves changes the student would wish to be made at their school. The average length of essays increased along with year level, from 98.2 words (first-year high school students) to 330.7 words (third-year high school/incoming college students).

2. Data Analysis

1) Identification of LBs

To address the first research question, LBs were identified in the three subcorpora, employing Salazar's (2014) method and using the concordance software AntConc (Anthony, 2022). Prior research on LBs has set different thresholds for what constitutes an LB, but the most common is a four-word-long sequence that occurs no less than 10 times per million words. For the present study, the frequency threshold for four-or-more word sequences was set at four times in the first-year corpus (about a half million words), two times in the second-year corpus (about a quarter million words), and three times in the third-year+ corpus (about 46,000 words). As Chen and Baker (2016) pointed out, to avoid inflating the numbers, LBs that quote the writing prompts from the dataset were removed. In addition, overlapping bundles within longer sequences were removed from the dataset.

The LBs were then categorized using structural and functional taxonomies developed in previous studies for the classification of LBs (e.g., Bychkovska & Lee, 2017;

Huang, 2015). The structural taxonomy involves identifying types of structural units: VP-based clausal bundles and NP- and PP-based phrasal bundles. The textual functions of the bundles were also categorized, according to their meaning in context. The three major categories of discourse functions in this study are: stance expressions (e.g., *I believe that the*), discourse organizers (e.g., *at the same time*), and referential expressions (e.g., *of the things that*).

2) Semantic Prosodies and Preferences

For the second part of the study, I compared the semantic preferences and prosodies of the LBs shared by all three groups in order to investigate how each group projects evaluative meaning when writing in response to identical writing prompts. For each LB, the analysis identified affective associations to determine its semantic prosody. This part of the analysis coded the LBs, following Xiao and McEnery's (2006) categorization, as positive, neutral, or negative. The category was determined by examining each LB in context, considering the adjoining structures to both left and right sides of the LB, employing the method devised by Y. Shin (2020).

The following illustration of the analysis of semantic prosody, from Y. Shin's (2020, p. 51) study, uses the LB *I would like to* as an example. This LB was identified in a corpus of L1-English university student writing. As shown in 1), the target bundle occurs in a main clause with a following complement. The preceding sentence was also taken into consideration for the analysis. The affect of the first sentence is negative: It describes a problem or difficulty (being unwanted) and overtly assesses it as negative ("not a good feeling"). In the second sentence, the LB is followed by the complement "change the views of my neighbors," whose semantic prosody is neutral. The labels in parenthesis at the end of the example indicate the semantic prosody that precedes (on the left) and follows (on the right) of the bundle: NEG for negative, NEU for neutral, and POS for positive.

1) It is not a good feeling to be unwanted in the place in which you reside. I would like to change the views of my neighbors. (NEG-NEU)

After the selected LBs' semantic prosodies were determined, their semantic preferences were identified by examining their co-occurring words in terms of meaning. In this way, the analysis identified semantic categories or lexical sets that recur frequently in association with specific LBs (e.g., Stubbs, 2001; Partington, 2004; Y. Shin, 2020).

IV. FINDINGS AND DISCUSSION

1. Structures and Textual Functions of LBs

This section addresses the first research question by identifying the LBs in the three corpora and analyzing them in terms of their structures and functions in context. The first-year corpus included 88 types of LBs; the second-year corpus, 93 types; and the third-year & higher corpus, 85 types. Four of the bundles are found in all three corpora, and these four comprise approximately 4% of the tokens of LBs in each corpus. Twenty LBs occur in both the first-year and the second-year corpora. Only two are shared between the second-year and the third-year & higher corpora, while four are shared between the first-year corpus and the third-year & higher corpus.

Figure 1 presents the main structures of LBs found in each of the three corpora. Notably, all three groups produced more VP-based bundles than other bundle types, but the proportion of VP-based bundles decreased by year (first-year: 90.9%, second-year: 70%, third-year & higher: 65.8%). Furthermore, the two higher year-level groups used more phrasal bundles than the first-year group; however, they favored different types of phrases, with more NP-based bundles in the second-year corpus and more PP-based bundles in the third-year & higher corpus.

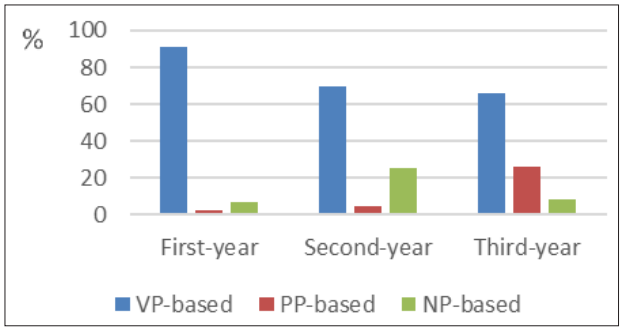


FIGURE 1 Distribution of Main Structural Categories

The three main structural types were then further subcategorized as in Table 2.

First, for all three groups, most of the VP-based bundles consist of a personal pronoun (almost always first-person) followed by a verb phrase, such as *I am not saying that* and *I would like to*; such usages sharply decreased by year (first-year: 63.2%, second-year: 46.2%, third-year+: 27.1%). Note that 80% of the 20 bundles shared between the first two year-level groups are in this subcategory, indicating that it is a feature that characterizes the lower year-level students' writing. Interestingly, most of these shared bundles also include smaller frequent sequences: *I want to* (8 bundles, e.g., *I really want to, why I want to, therefore I want to*) and *I think* (4 bundles, e.g., *I think this is, I think we should*). This finding is consistent with a claim in the literature that low proficiency English learners tend to latch onto certain sets of lexical frames when producing multiword sequences (e.g., Y. Shin et al., 2021). On the other hand, the structure of VP-based bundles favored by the third-year+ group involves anticipatory-*it*, along with verb/adjective phrases, such as *it is true that* and *it is obvious that*. This type of bundle accounts for about 2% of the first- and second-year groups' LBs, compared to 7.7% for the third-year+ group.

With respect to PP-based bundles, the third-year+ stu-

TABLE 2
Distribution of Structural Subcategories

Category	Subcategory	First-year		Second-year		Third-year+	
		Types	Tokens	Types	Tokens	Types	Tokens
VP-based	Personal pronoun + verb phrase (e.g., *I believe that the*, *I am going to*)	55.7% (49)	63.2% (520)	35.5% (33)	46.2% (197)	25.9% (22)	27.1% (95)
	(VP) + *that*-clause fragment (e.g., *that I want to*, *that we have to*)	1.1% (1)	1% (8)	4.3% (4)	6.6% (28)	1.2% (1)	1.7% (6)
	(Verb/adjective) *to*-clause fragment (e.g., *to learn how to*, *to go to the*)	2.3% (2)	8.5% (70)	4.3% (4)	4.2% (18)	7% (6)	6.3% (22)
	Existential-*there*construction (e.g., *there is only one*)	2.3% (2)	1.2% (10)	5.4% (5)	4.2% (18)	4.7% (4)	4.5% (16)
	Anticipatory *it*+ VP/AP (e.g., *it is better than*, *it is true that*)	3.4% (3)	2.5% (21)	3.2% (3)	2.1% (9)	9.4% (8)	7.7% (27)
	Copula *be*+ noun phrase/adjective phrase (e.g., *this is why I*)	15.9% (14)	11.5% (95)	9.7% (9)	7.5% (32)	12.9% (11)	12.2% (43)
	(Verb phrase) + active verb (e.g., *sign up for the*)	5.7% (5)	3.8% (31)	5.4% (5)	4.9% (21)	5.9% (5)	6.3% (22)
PP-based	PP with embedded *of*-phrase (e.g., *in the case of*)	0% (0)	0% (0)	0% (0)	0% (0)	1.2% (1)	0.8% (3)
	Other PP fragment (e.g., *on the other hand*, *in my high school*)	3.4% (3)	2.4% (20)	5.4% (5)	4.9% (21)	22.3% (19)	25.1% (88)
NP-based	NP with *of*-phrase fragment (e.g., *a high voice of*)	0% (0) 0	0% (0)	3.2% (3)	2.1% (9)	2.3% (2)	2% (7)
	NP with other postmodifier fragment (e.g., *the best way to*)	5.7% (5)	3.1% (26)	11.8% (11)	10.3% (44)	5.9% (5)	5.1% (18)
	Other noun phrase (e.g., *a lot of information*, *a lot of homework*)	4.5% (4)	2.7% (22)	11.8% (11)	6.8% (29)	1.2% (1)	1.1% (4)
Total		100% (88)	100% (823)	100% (93)	100% (426)	100% (85)	100% (351)

Note. The numbers in parentheses refer to occurrences in the corpus.

dents used them the most (25.1% of all their bundles; first-year: 2.4%, second-year: 4.9%). Nevertheless, PPs with embedded *of*-phrase fragments, which are indicative of proficient academic writers according to previous literature (e.g., Bychkovska & Lee, 2017; Salazar, 2014), were scarce. The first two adjacent groups never produced such a structure; the third-year & higher group used very few (about 1% of their LBs, e.g., *in the case of*). Most of the PP-based bundles found in this study were not in fact the same as those that have been reported to be characteristic of academic prose. Instead, they mostly consisted of idiomatic expressions, including *for a long time* and *in the real world*. Similarly, the NP-based bundles found in this study are among those often subcategorized as "other noun phrase" in prior research, indicating that these bundles are atypical of academic writing. Many of them involve colloquial quantifiers such as *a lot of* and *lots of* (e.g., *a lot of knowledges*), which have been labeled "learner bundles" in previous studies (e.g., Bychkovska & Lee, 2017; Chen & Baker, 2016; Huang, 2015; Y. Shin, 2019). Many of the NP-based bundles also included grammatical errors.

Figure 2 shows the discourse functions of the bundles in each of the three corpora. As seen in the figure, stance expression bundles comprise the largest proportion in the first-year corpus (about 70% of all bundles) whereas the three main function types (i.e., stance expressions, referential expressions, and discourse organizers) are relatively evenly distributed in the two higher year-level corpora.

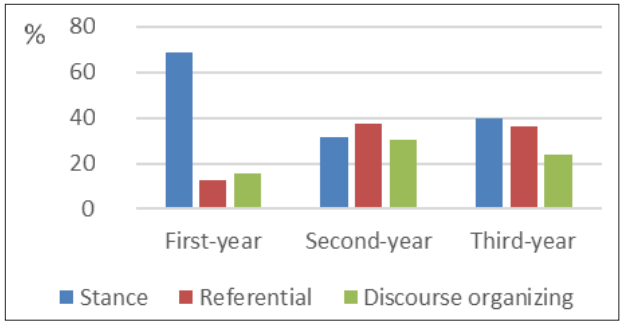


FIGURE 2 Distribution of Main Functional Categories

Table 3 shows functional subcategories of bundles found in each of the three corpora. The first-year group's use of stance expression bundles is notably higher, in line with prior work that has suggested that overuse of stance expressions is a feature of the writing of apprentice writers and/or language learners, in contrast to the writing of expert/native academic writers (e.g., Chen & Baker, 2010, 2016; Y. Shin, 2021; Staples et al., 2013).

With respect to discourse organizing bundles, the two higher year-level groups produced them at similar rates (both over 30%), but with different preferences for specific bundles or bundle structures. The second-year group was more likely to produce similar sets of bundles with varied conjunctive

adverbials (e.g., *first*, *second*, and *therefore*) in the initial position, for example, *first I want to*, *second I want to*, and *therefore I want to*, while the third-year group preferred *on the other hand* (the most frequent bundle in the third-year corpus) and *when it comes to*. These bundles were not found in the first- and second-year corpora.

As for referential bundles, the second-year corpus contained a notably high number of them (about 25% of all bundles), which appears to resemble a norm of academic prose (e.g., Yang et al., 2022). However, close scrutiny of the actual bundles revealed that they were not the same referential bundles documented in the literature as characteristic of proficient academic prose (e.g., Pan et al., 2016; Salazar, 2014; D. Shin & Y. Shin, 2020). The referential bundles used by the second-year group include *have a lot of*, *there are many things that*, and *a lot of homework* (categorized as quantity specification) and *students who want to* and *thing in our school* (categorized as identification/focus); most of these are considered colloquial expressions characteristic of face-to-face conversation.

2. Semantic Prosodies and Preferences

This section explores the uses of the shared bundles (those found in two or more corpora) in terms of their semantic prosodies and semantic preferences in order to examine how the same sets of bundles were used under the same writing topics by different groups. Among the shared bundles, three were chosen based on frequency: *so I want to*, shared by the first- and second-year corpora; *will be able to*, shared by the first-year and third-year & higher corpora; and *I would like to*, shared by all three corpora. (The two LBs shared by the two higher year-level corpora occur less than five times in each corpus and were excluded from this analysis due to infrequency.)

1) *"so I want to"*

In both the first- and second-year corpora, this bundle always occurs in a main clause, mostly in the sentence-initial position. The semantic analysis of *so I want to* therefore considered the semantic prosodies of the preceding sen-

TABLE 3
Distribution of Functional Subcategories

Category	Subcategory	First-year		Second-year		Third-year+	
		Types	Tokens	Types	Tokens	Types	Tokens
Stance expressions	Epistemic (e.g., *I think this is*, *due to the fact*)	17% (15)	12.6% (104)	10.7% (10)	8.7% (37)	8.2% (7)	7.4% (26)
	Attitudinal/Modality (e.g., *I really want to*, *which is more important*)	45.4% (40)	56.2% (463)	24.7% (23)	22.8% (97)	32.9% (28)	32.2% (113)
Discourse organizers	Topic introduction (e.g., *is the next things*, *when it comes to*)	2.3% (2)	3.5% (29)	3.2% (3)	5.9% (25)	2.3% (2)	4.3% (15)
	Topic elaboration/clarification (e.g., *it doesn't mean that*)	13.6% (12)	12.3% (101)	18.3% (17)	24.6% (105)	20% (17)	19.6% (69)
Referential expressions	Identification/Focus (e.g., *is one of the*, *there is one thing*)	4.5% (4)	2.5% (21)	10.7% (10)	11.7% (50)	3.5% (3)	3.7% (13)
	Framing attributes (e.g., *according to my experience*, *in the case of*)	0% (0)	0% (0)	6.4% (6)	4.7% (20)	5.9% (5)	5.4% (19)
	Quantity specification (e.g., *contains a lot of*, *are larger than now*)	6.8% (6)	3.8% (31)	16.1% (15)	12.2% (52)	8.2% (7)	7.4% (26)
	Place/Time/Text-deixis (e.g., *a place where we*, *when I was young*)	7.9% (7)	6.4% (53)	8.6% (8)	8.7% (37)	18.8% (16)	20% (70)
Conversational function	Politeness (e.g., *thank you very much*, *thank you for reading*)	2.3% (2)	2.5% (21)	1% (1)	0.7% (3)	0% (0)	0% (0)
Total		100% (88)	100% (823)	100% (93)	100% (426)	100% (85)	100% (351)

Note. The numbers in parentheses refer to occurrences in the corpus.

TABLE 4
Semantic Prosodies that Precede (to the Left) and Follow (to the Right) of *so I want to* in the First-year and Second-year Corpora

Corpus	Left	Negative			Neutral			Positive		
	Right	NEG	NEU	POS	NEG	NEU	POS	NEG	NEU	POS
First-year (n = 106)	Token	1	72	0	0	28	0	0	5	0
	%	0.9%	67.9%	0%	0%	26.4%	0%	0%	4.7%	0%
Second-year (n = 40)	Token	0	27	0	0	12	0	0	1	0
	%	0%	67.5%	0%	0%	30%	0%	0%	2.5%	0%

tences and the following complements (i.e., *to*-infinitives). Table 4 shows the findings for the two corpora.

As seen in the table, the bundle's complement (to the right of the LB) is almost always neutral in both corpora. The preceding sentence is dominantly negative (over 67% for both groups), sometimes neutral, and occasionally positive.

The following examples show uses of *so I want to* found in the two corpora in essays written in response to the same topic (something the writer would wish to be changed at their school). In both examples, the student writers first present a problem (negative), and then use the bundle to introduce what they want to be changed (neutral). The shared bundle is indicated in bold in the examples.

2) In other words, playing basketball outside has many physically serious problems. **So I want to** make it. (First-year, NEG-NEU)

3) Especially some part of gym is broken and very dirty. **So I want to** change gym center, if I can. (Second-year, NEG-NEU)

Four main semantic categories recur in essays written in response to this prompt in both corpora, although their frequencies differ: school facilities, including classroom environments (first-year: 29%, second-year: 60.5%); the educational system, such as curricula and teaching methods (first-year: 28.6%, second-year: 14%); uniforms (first-year: 16.2%, second-year: 6.2%); and school rules (first-year: 15%, second-year: 8.1%).

Examples 4) and 5) demonstrate how this LB occurs with the same semantic category (i.e., educational system) in each corpus. In both cases, as in the previous examples, the first sentence, to the left of the LB, describes a problem (not enough time, not a good system), and the LB is used to introduce the change the writer wants to make in the second sentence, to the right of the LB. Thus, the sequence of semantic prosodies is again NEG-NEU.

4) Because high school's test is to keep from student focusing on themselves studying. I need for a long time that I was focusing on my study, but Korean education process would not give the 'times' to student. **So I want to** change the education process. (First-year, NEG-NEU)

5) I want to learn physics but life science is not, I think students have to choose subject that they want, our school's system is not good for students. **So I want to** change science focusing classroom system. (Second-year, NEG-NEU)

2) *"will be able to"*

As Table 5 shows, while the first-year and third-year+ corpora both contain the bundle *will be able to*, the two groups display marked differences in their use of it. The analysis of semantic prosodies shows that both student groups predominantly project positive prosody after the bundle, but what comes before the bundle greatly differs between the two groups.

The following examples show the different patterns of the bundles in each corpus. As shown in 6), the first-year student displayed positive prosody before and after the bundle, listing advantages of having songs played during lunch time at school. In 7), however, the third-year student used the bundle in a *that*-clause with a neutral evaluation, preceded by a negative main clause ("there is no guarantee"). This pattern of NEG-NEU was found to be the second most frequent prosody in the third-year+ corpus, and did not occur in the first-year corpus.

6) If songs are played in full time, it can be not only entertain for students but also energy of school's lunch time and we **will be able to** have happy lunch time. (First-year, POS-POS)

7) But there is no guarantee that he **will be able to** follow those instructions, let alone remember them. (Third-year+, NEG-NEU)

Both groups, however, tended to use the same affective prosody pattern of NEU-POS when the bundle functions to state the benefits of the change they wished to make. In 8), the first-year student explains the wish for a 3D printer to the left of the LB, and then uses the LB to introduce the merits of this idea. Similarly, in 8) the third-year+ writer stated the school rule that he or she wished for before the bundle, and then used the bundle to introduce the description of the benefits of the rule.

8) Because of these reasons, I want to place a 3D printer in the school. It **will be able to** give students dream and hopes. (First-year, NEU-POS)

TABLE 5

Semantic Prosodies that Precede (to the Left) and Follow (to the Right) of *will be able to* in the First-year and Third-year+ Corpora

Corpus	Left	Negative			Neutral			Positive		
	Right	NEG	NEU	POS	NEG	NEU	POS	NEG	NEU	POS
First-year	Token	0	0	0	0	0	5	0	1	6
(*n* = 12)	%	0%	0%	0%	0%	0%	41.7%	0%	8.3%	50%
Third-year+	Token	0	2	1	0	1	5	0	0	0
(*n* = 9)	%	0%	22.2%	11.1%	0%	11.1%	55.5%	0%	0%	0%

9) I'll restrict outsiders from entering the festival stage and allow only students from our school to enter. It **will be able to** watch the stage more comfortably than before. (Third-year+, NEU-POS)

In sum, the first-year corpus exhibited one recurrent semantic preference, POS-POS, collocated with this bundle, as in 6) above, often with preceding sentences about being happy and relaxed (33.3%), while no recurrent pattern was found in the third-year & higher corpus.

3) *"I would like to"*

The bundle *I would like to*, shared by the three corpora, shows similar semantic patterns in the writing of the two adjacent lower year-level groups, and a different pattern in the writing of the third-year+ group.

As Table 6 presents, the bundle showed a tendency toward the same semantic meaning between the first- and second-year corpora. In these two adjacent lower year-level corpora, the bundle was most frequently preceded by a sentence with neutral affect, and followed by a neutral complement as well. On the other hand, in the third-year & higher corpus, half of the instances of this bundle are followed by a positive-affect complement, while such semantic prosody never occurs in either the first-year or the second-year corpus.

The following examples illustrate the most frequently recurring affective meaning pattern in the LBs' co-occurring text in each of the three corpora: first-year, in 10), and second-year, in 11), show neutral prosody both before and after the bundle; third-year & higher, in 12), shows negative prosody preceding and positive prosody following the bundle.

10) Then, what change could be made to make our school better? In order to make our school a better place, **I would like to** suggest doing a debate class once a week. (First-year, NEU-NEU)

11) To summarize, the road (site) can handle our hometown's image. For these reason, **I would like to** change the road (site) in our hometown. (Second-year, NEU-NEU)

12) However, it was difficult to spend time to gather other than class or lunch time because students lived far away from school so that it was hard to find a place to meet up together and also the school didn't allow much club or activities after school. So **I would like to** make a change to the school policy allowing students to create their own clubs and activities after school, so that they can hang out more and spend time together to get to know about each other. (Third-year+, NEG-POS)

As for semantic preference, one obvious semantic set found across the three corpora involves direct responses to the essay prompt. All three groups were highly likely to address the prompt topic directly at the beginning of the whole essay, as the thesis statement. While the bundle *I would like to* is considered typical of colloquial spoken language (e.g., Y. Shin, 2019; Staples et al., 2013), for this population of Korean EFL students, it functions to state their main argument, regardless of their English proficiency. Many of the nouns in the complements of the bundle (e.g., "school policy," "food in cafeteria," "school uniform," "the way of taking tests") directly address the essay prompt. Furthermore, it was not uncommon, in all three corpora, to find the bundle preceded by an *if*-clause that repeated the essay topic.

13) If I could change my school life, **I would like to** change food system, unless we change this problem we will hurt to each other. (First-year, NEU-NEG)

14) If I could make one important change in a school that I'm attending. **I would like to** abolish test. (Second-year, NEU-NEU)

15) If I could change one specific thing in my high school I attended, **I would like to** make practical subjects non-mandatory. (Third-year+, NEU-NEU)

V. CONCLUSION

This study investigated the uses of LBs by young Korean EFL students of different school year-levels in

TABLE 6
Semantic Prosodies that Precede (to the Left) and Follow (to the Right) of *I would like to* in the Three Corpora

Corpus	Left	Negative			Neutral			Positive		
	Right	NEG	NEU	POS	NEG	NEU	POS	NEG	NEU	POS
First-year (n = 25)	Token	0	7	0	1	13	0	0	5	0
	%	0%	28%	0%	4%	52%	0%	0%	20%	0%
Second-year (n = 19)	Token	0	1	0	1	15	0	0	1	0
	%	0%	5.3%	0%	5.3%	78.9%	0%	0%	5.3%	0%
Third-year+ (n = 6)	Token	0	0	3	0	2	0	0	1	0
	%	0%	0%	50%	0%	33.3%	0%	0%	16.7%	0%

their essay responses to the same writing topics. The first part of the study showed clear differences between the first-year high school student group and the two higher year-level student groups – the former produced LBs characteristic of face-to-face conversation in structure while the latter groups used more of the phrasal bundles considered characteristic of proficient academic writers.

On the other hand, the second part of the study, which examined the semantic prosodies and preferences of shared bundles (those found in more than two corpora) exhibited different pictures of bundle usage by group. That is, even when they used the same bundle in response to the same prompt, the different groups often employed the bundle in different ways. Thus, while the two higher year-level groups showed similar patterns of bundle usage in terms of the bundles' internal structures, they showed markedly different usage in terms of semantic properties, which means that they projected different evaluative meanings via bundles in context.

The results of the current research provide immediate classroom practice implications for teachers of EFL writers as well as novice academic writers. The types of LBs shared by the students across different grade levels and those unique to each group that this research identified (See Appendix) could serve as a useful resource in teaching structures common in English academic writing to these young EFL populations. Moreover, the second question asked to what extent high school student writers project semantic properties via LBs in their argumentation. This finding of Korean EFL writers' avoidance of negative phrasing echoes the results of Y. Shin's (2022) comparison of LB use by native and nonnative college student writers; the difference in amounts of negative/positive prosody between native and nonnative writers of English, even when they are writing on the same topic, is striking. In both studies, the learners almost never presented negatively valenced examples, even as counterarguments. Meanwhile, native writers appear to use negative phrasing in argumentative essays freely, suggesting that presenting negative counterarguments could be considered a characteristic of argumentation. The pedagogical implication is that instructors might find it useful to remind nonnative English writers to present both positive and negative ideas to support their thesis. On the other hand, it is possible that novice native writers overuse negative phrasing, suggesting that they could also benefit by being guided to balance their arguments. The findings on how EFL writers project affective meanings should be useful for guiding learners to consider how they can present positive and negative (counter) examples to lead to and support their arguments.

Overall, the study provides educators information on the structures preferred by Korean school students, enabling the development of pedagogical materials that address the use of these structure types and their co-occurrence with register-appropriate linguistic complexity features.

One major limitation of the study is the small size of the corpus used for the analysis. As mentioned earlier, this limitation was in part due to the pandemic situation, in which the in-person visits to Korean high schools were not completed as planned, which hindered the collection of the data of student essays. While the findings are still meaningful, as they revealed clear patterns of structural and semantic tendencies from this first attempt to build a developmental corpus of secondary school student essays, there is an evident need to augment the size of the dataset for future studies. Once the corpus is large enough (i.e., approximately a million words for each year-level), it will be possible to conduct more robust corpus-based studies on this under-researched population (i.e., EFL secondary school students).

Another limitation, as pointed out by one of the reviewers, involves the data collection. While the high school essays came from different schools, the incoming college student essays were from one Korean university. It is thus possible that the college student essays are less representative of students' school level (i.e., year) than of the particular university setting, indicating the need to collect data from more varied university environments. Moreover, the present study did not group learners by their English proficiency level, but by school year; a future study should assess English proficiency levels to provide a more concrete picture of developmental sequences of bundle usage.

On a final note, this study lays the groundwork for an important line of research that can be expected to yield valuable pedagogical results. Studies comparing parallel corpus data controlled for register and writing prompt are rare, and little research has compared secondary school student essays on identical topics among different English proficiency level groups. The corpus used in this research should be especially useful for achieving better understanding of this under-researched population. As noted above, while many comparative corpus-based studies have targeted student writers at the university level, a dearth of research has looked at student writers at earlier stages. Abundant research has shown that factors such as proficiency, genre, topic, and time constraints directly affect written production. Nevertheless, surprisingly, extant corpus studies on writing development have largely failed to employ well-matched corpora. For future research, this developmental corpus data could be used to support a wide range of research on how different groups of EFL learners use English and on the development of L2 writing ability over time, as well as to inform English writing pedagogy to better facilitate learners on their journey from novice writers to proficient writers of academic English.

REFERENCES

Anthony, L. (2022). AntConc (Version 4.0.6) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

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

Bychkovska, T., & Lee, J. (2017). At the same time: Lexical

bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes, 30*, 38-52.

Chen, Y., & Baker, P. (2016). Investigating critical discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics, 37*(6), 849-880.

Cortes, V., & Hardy, J. (2013). Analyzing the semantic prosody and semantic preference of lexical bundles. In D. Belcher & G. Nelson (Eds.), *Critical and corpus-based approaches to intercultural rhetoric* (pp. 180-201). University of Michigan Press.

Huang, K. (2015). More does not mean better: Frequency and accuracy analysis of lexical bundles in Chinese EFL learners' essay writing, *System, 53*, 13-23.

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

Kwon, Ye-Eun. (2021). An analysis on lexical bundles of college students' speaking scripts: Based on a self-introduction task. *Modern English Education, 22*(4), 79-89.

Moon, R. (1992). Textual aspects of fixed expressions in learners' dictionaries. In P. J. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 13-26). Claredon Press.

Morley, J., & Partington, A. (2009). A few frequently asked questions about semantic – or evaluative – prosody. *International Journal of Corpus Linguistics, 14*(2), 139-158.

Northbrook, J., & Conklin, K. (2019a). Is what you put in what you get out? – Textbook-derived lexical bundle processing in beginner English learners. *Applied Linguistics, 40*(5), 816-833.

Northbrook, J., & Conklin, K. (2019b). "What are you talking about?" An analysis of lexical bundles in Japanese junior high school textbooks. *International Journal of Corpus Linguistics, 23*(3), 311-334.

Oster, U. (2010). Using corpus methodology for semantic and pragmatic analyses: What can corpora tell us about the linguistic expression of emotions? *Cognitive Linguistics, 21*(4), 727-763.

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

Paquot, M. (2017). L1 frequency in foreign language acquisition: Recurrent word combinations in French and Spanish EFL learner writing. *Second Language Research, 33*(1), 13-32.

Park, Eunjeong. (2019). The effectiveness of corpus-aided instruction to improve second language college students' academic writing. *Modern English Education, 20*(3), 15-25.

Partington, A. (2004). "Utterly content in each other's company": Semantic prosody and semantic preference. *International Journal of Corpus Linguistics, 9*(1), 131-156.

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

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching*. John Benjamins.

Shin, Do Hwan, & Shin, Yu Kyoung. (2020). Retaliated with tariffs on: A corpus-based analysis of lexical bundles in TED talks and BBC news in Global Business issues. *Modern English Education, 21*(2), 71-84.

Shin, Yu Kyoung. (2018). The construction of English lexical bundles by native and nonnative freshman university students. *English Teaching, 73*(3), 113-137.

Shin, Yu Kyoung. (2019). Do native writers always have a head start over nonnative writers? The use of lexical bundles in college students' essays. *Journal of English for Academic Purposes, 40*, 1-14.

Shin, Yu Kyoung. (2020). Evaluative prosody and semantic preference: Extending the analysis of recurrent multiword sequences. *English for Specific Purposes, 59*, 42-58.

Shin, Yu Kyoung, Cortes, V., & Yoo, Isaiah WonHo. (2018). Using lexical bundles as a tool to analyze definite article use in L2 academic writing: An exploratory study. *Journal of Second Language Writing, 39*, 29-41.

Shin, Yu Kyoung, & Kim, YouJin. (2017). Using lexical bundles to teach articles to L2 English learners of different proficiencies. *System, 69*, 79-91.

Shin, Yu Kyoung, Koo, Yeri, Park, Yerin, & Seo, Donju. (2021). Argument-related recurrent sequences: A corpus-based study targeting L2 learners of English. *Secondary English Education, 14*(4), 52-89.

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

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes, 12*(3), 214-225.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Blackwell.

Wei, Y., & Lei, L. (2011). Lexical bundles in the academic writing of advanced Chinese EFL learners. *RELC Journal, 42*(2), 155-166.

Xiao, R., & McEnery, T. (2006). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics, 27*(1), 103-129.

Yang, L., Nikitina, L., & Riget, P. (2022). Development of syntactic complexity in Chinese university students' L2 argumentative writing. *Journal of English for Academic Purposes*, *44*, 101099. doi: 10.1016/j.jeap.2022.101099

Yoon, Choongil, & Choi, Ji-Myoung. (2015). Lexical bundles in Korean university students' EFL composition: A comparative study of register and use. *Modern English Education, 16*(3), 47-69.

APPENDIX

Distribution of Lexical Bundles Used by the Three Groups

First-year high school (88 types, 823 tokens)		Second-year high school (93 types, 426 tokens)		Third-year high school/ Incoming college (85 types, 351 tokens)	
so I want to	106	so I want to	40	on the other hand	13
to change our school	60	I would like to	19	I was able to	10
and I want to	31	that I want to	13	when it comes to	9
I would like to	25	thing I want to	11	will be able to	9
I think it is	20	when we go to	9	for a long time	8
I think our school	18	I don't want to	9	there are a lot of	7
second I want to	17	are larger than now	8	I would like to	6
I hope to change	16	for this reason I	7	I was in high school	6
first I want to	15	that we have to	7	when I was young	6
I don't want to	14	we don't have to	7	when I went to	6
if I can make	13	first I want to	6	in my high school	6
it is hard to	13	have a lot of	6	that I want to	6
school uniform is too	13	there is only one	6	I would want to	6
will be able to	12	to go to school	6	at the same time	5
I will change the	11	when I go to	6	in our daily lives	5
why I want to	11	the rule that we	6	in my case I	5
hello my name is	11	will be improved the	6	due to the fact	5
to go to school	10	because we have to	5	it is true that	5
when I go to	10	reason I want to	5	which is more important	5
I like my school	10	so if I could	5	because they do not	4
thank you for reading	10	spend too much time	5	so if I could	4
so we have to	9	thing that I want	5	because of this reason	4
therefore I want to	9	I think it is	5	for example when a	4
we go to school	9	I think that the	5	for example when i	4
I can make one	9	to focus on studying	5	however in my opinion	4
for these reasons I	8	that we want to	5	a lot of information	4
for a long time	8	what I want to	5	I am going to	4
I want to be	8	our school's education system	4	I went to the	4
I want to need	8	and I want to	4	when I was a	4
I want to turn	8	but I want to	4	things that we can	4
that I want to	8	second I want to	4	from the real life	4
because I don't like	7	therefore I want to	4	the contents of the	4
a lot of time	7	for these reasons I	4	these reasons I think	4
the time when we	7	test score is the	4	to the fact that	4
I think we should	7	students who want to	4	reasons why I think	4
So I think that	7	the other one so	4	have a chance to	4
from experience is better	7	thing in our school	4	not be able to	4
is too small to	7	to change the way	4	to get to know about	4
so if I could	6	sign up for the	4	to learn how to	4
contains a lot of	6	up for the program	4	to make a change to	4
a lot of knowledges	6	I don't know why	4	was able to get	4
rules that students must	6	I think this is	4	I believe that the	4
I think I want	6	the point of view	4	it is hard to	4
I think that the	6	how bad and sorrowful an examination system is	4	the best way to	4
so I think the	6	I love my school	4	but if I have	3

First-year high school (88 types, 823 tokens)		Second-year high school (93 types, 426 tokens)		Third-year high school/ Incoming college (85 types, 351 tokens)	
and we have to	6	I really want to	4	I am not saying that	3
don't like our school	6	we need time to	4	if I had the	3
I'm satisfied with my	6	but if I could	3	if you want to	3
I really want to	6	if I want to	3	it doesn't mean that	3
I want our school	6	if you want to	3	these are the reasons	3
is more than experiences	6	why I want to	3	this is why I	3
is not good for	6	first of all I	3	first of all I	3
is the most important	6	are many things that	3	in my opinion I	3
we will be able	6	there are lots of	3	there are many ways	3
but if I could	5	there are many things	3	there are so many	3
I was in middle school	5	there are so many	3	there were a lot of	3
but there is one	5	there are too many	3	were a lot of	3
there is one thing	5	is too small to	3	a wide range of	3
one thing that I	5	to go to the	3	as time went by	3
and I think the	5	we go up to	3	as soon as I	3
but I think that	5	too much time in	3	all over the world	3
change the rule of	5	is only one way	3	at some point you	3
I don't like this	5	is one of the	3	at the time I	3
I love my school	5	that they want to	3	for the first time	3
is too uncomfortable to	5	according to my experience	3	the time when I	3
what I want to	5	the way of making	3	the problem is that	3
is the next things	4	I also think that	3	how to deal with	3
finally I want to	4	I can't understand this	3	in the case of	3
first we have to	4	I know it's hard to	3	I think it is	3
the reason why I	4	I think that it	3	know what to do	3
both of them are	4	I think we should	3	so I think the	3
a lot of money	4	be able to eat	3	I did not have	3
lot of time to	4	is much better than	3	can be obtained from	3
in the real world	4	it is better than	3	can get good grades	3
I think I want to	4	it should be changed	3	I learned a lot	3
I think it will	4	of course it is	3	to come up with	3
I think that it	4	students have to study	3	are more important than	3
I think this is	4	all I want to	3	I wish I could	3
so I think our	4	thank you for reading	3	it is difficult to	3
so I think we	4	before we go into the	2	it is obvious that	3
and it is too	4	a good thing but	2	it was difficult to	3
can give us chances to	4	a lot of books	2	it was hard to	3
want to need a	4	a lot of homework	2	it was really hard	3
I wish I could	4	a lot of people	2	will not be able to	3
is too short to	4	a lot of stressed	2	a chance to try	3
isn't obvious and the	4	a lot of time	2		
it is not good	4	a place where we	2		
we don't need to	4	a high voice of	2		
		a bad thing at all	2		
		a big obstacle for	2		
		a few differences from	2		
		a limited area the	2		
		best choice for everyone	2		

Note. LBs shared by all three corpora (n = 4) are indicated in gray; LBs shared by first-year and second-year corpora (n = 20) are in blue; LBs shared by second-year and third-year+ corpora (n = 2) are in purple; LBs shared by first-year and third-year+ corpora (n = 4) are in red.