*Modern English Education*, Vol. 24, (2023) pp. xx-xx

****doi: https://doi.org/10.18095/meeso.2023.24.1.xxx



**Modern English Education**

ISSN: 1598-0782 / e-ISSN: 2586-6141

http://journal.meeso.or.kr

**Semantic Properties of Lexical Bundles Used by Young Adult Korean EFL Students at Different Grade Levels** (Times New Roman/바탕체)

**Kildong Hong** (Hankook University)

**Younghee Kim** (Hankook University)

Received: 28 April 2023

Revised: 25 May 2023

Accepted: 1 June 2023

**Keywords**

Korean EFL student, semantic prosody, lexical bundle

한국인 영어학습자, 의미적 운율, 말뭉치

**Kildong Hong** (First author)

PhD Candidate

Department of English Language Education

Hankook University

kdhong@hankook.ac.kr

ISNI: 0000 0000 0000 0000

**Younghee Kim** (Corresponding author)

Professor

Department of English Language Education

Hankook University

younhk@hankook.ac.kr

ISNI: 0000 0000 0000 0000

\*This work was supported by 2022 Social Sciences Research Grant to Kildong Hong.

**Hong, Kildong, & Kim, Younghee. (2023). Semantic properties of lexical bundles used by young adult Korean EFL students at different grade levels. *Modern English Education, 24*, xx-xx.**

**Abstract**

Meta-analytic reviews collect available empirical studies on a speciﬁed domain and calculate the average effect of a factor. Educators as well as researchers exploring a new domain of inquiry may rely on the conclusions from meta-analytic reviews rather than reading multiple primary studies. This article calls for caution in this regard because the outcome of a meta-analysis is determined by how effect sizes are calculated, how factors are deﬁned, and how studies are selected for inclusion. Three recently published meta-analyses are reexamined to illustrate these issues. The ﬁrst illustrates the risk of conﬂating effect sizes from studies with different design features; the second illustrates problems with delineating the variable of interest, with implications for cause-effect relations; and the third illustrates the challenge of determining the eligibility of candidate studies. Replication attempts yield outcomes that differ from the three original meta-analyses, suggesting also that conclusions drawn from meta-analyses need to be interpreted cautiously. the second illustrates problems with delineating the variable of interest, with implications for cause-effect relations; and the third illustrates the challenge of determining the eligibility of candidate studies. Replication attempts yield outcomes that differ from the three original meta-analyses, suggesting also that conclusions drawn from meta-analyses need to be interpreted cautiously.

**INTRODUCTION** (First Subheading, Arial Bold/맑은 고딕 진하게)

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 dis- course 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 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.

**LITERATURE REVIEW**

Lexical Bundles (Second Subheading, Arial/맑은 고딕)

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., Kwon, 2021; Park, 2019; Salazar, 2014; Shin et al., 2018), and demonstrated that the ability to use them competently is a key element of academic skills (e.g., Bychkovska & Lee, 2017; Lee et al., 2020).

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).

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., Kwon, 2021; Yoon & Choi, 2015) and professional academics (e.g., Salazar, 2014). The absence of systematically com- piled 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.

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; Shin, 2020; Sinclair, 2004; Stubbs, 2001).

The 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).

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

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

**METHOD**

Corpus Data

The study uses a corpus of English argumentative es- says 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 short- fall was partly made up by including essays on the same topics written by incoming college students, which had al- ready 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).

Data Analysis

*Identification of LBs (Third Subheading, Arial/맑은 고딕 )*

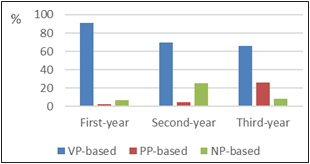
To address the first research question, LBs were identified in the three subcorpora, employing Salazar’s (2014) method and using the concordance software AntConc (Anthony, 2022). Prior research on LBs has set different thresholds for what constitutes an LB, but the most common is a four-word-long sequence 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.

**FINDINGS AND DISCUSSION**

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; how- ever, 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 and 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 an- ticipatory-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+ students 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 and 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; Shin, 2019). Many of the NP-based bundles also included grammatical errors.

**CONCLUSION**

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

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.

Hadjerrouit, S. (2020, September). *Exploring the affordances of Numbas for mathematical learning: A case study* [Paper presentation]. DRUM 2020, Tunis, Tunisia. https://hal.science/hal-03113967v1/document

Halliday, M. A. K., & Hasan, R. (2014). *Cohesion in English*. Routledge.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfield, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Newbury House.

Kim, Hyekyeng. (2015). *The effects of pragmatic instruction on Korean university students with different language proficiency levels*. [Unpublished doctoral dissertation]. Hankuk University of Foreign Studies.

Kong, Eun Jung., & Yoon, In Hee. (2013). L2 proficiency effect on the acoustic cue-weighting pattern by Korean L2 learners of English: Production and perception of English stops. *Journal of the Korean Society of Speech Sciences*, *5*(4), 81–90.

Lee, Eun-Hee. (2014). An analysis of Korean EFL university learners’ use of cohesion in writing with Coh-Metrix. *The Journal of Foreign Studies*, *27*, 195–216.

Lim, Hyojung. (2023). Exploring the development of second language (L2) vocabulary in speaking. *Studies in Foreign Language Education, 37*(1), 145–163.

Machicao, J. C. (2019). Higher education challenge characterization to implement automated essay scoring model for universities with a current traditional learning evaluation system. In Á. Rocha, C. Ferrás & M. Paredes (Eds.), *Information technology and systems: Advances in intelligent systems and computing* (pp. 835–844). Springer.

Malmcrona, H. (2020). *Variation in assessment: A Coh-Metrix analysis of evaluation of written English in the Swedish upper secondary school*. [Unpublished master’s thesis]. Stockholm University.

Peters, E. (2019). Factors affecting the learning of single-word items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 125–142). Routledge.

Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In M. Lapata, H. & Tou Ng (Eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 86–195). Association for Computational Linguistics.

Appendix

Materials for consciousness-raising activity

Sample 1

|  |
| --- |
| Dear Mr. Ko  Hello this is Uhyeong Jeong.  I am writing this email to notice I can't attend the class on March 23. Because I broke my leg in a car accident.  So, could u accept my attendance at class on March 23?  Bye |