English Teaching, Vol. 73, No. 3, Autumn 2018
DOI: 10.15858/engtea.73.3.201809.115

The Construction of English Lexical Bundles in Context by Native and Nonnative Freshman University Students

Yu Kyoung Shin
(Hallym University)

Shin, Yu Kyoung. (2018). The construction of English lexical bundles in context by native and nonnative freshman university students. *English Teaching, 73*(3), 115-139.

This study investigates how newcomers to the university setting integrate lexical bundles (LBs)—frequently recurring word sequences—into their writing by analyzing the bundles' *syntactic roles* (i.e., relations to larger structures). Previous studies have considered phrases and clauses as the main internal structures of LBs; however, these boundaries might not always be clear because such fragmented sequences do not stand alone, but are embedded in larger structures. The current study addresses this gap by investigating the syntactic roles of LBs identified in native and nonnative corpora of English argumentative essays (approximately 1400 essays, comprising half a million words each) written in response to identical writing prompts. The results show that the two language groups display generally similar patterns of using LBs due to their status as novice academic writers, but at the same time, their patterns do diverge to some extent, with some uses unique to or more common to each group. The extent to which different population groups use LBs in relation to the syntactic complexity typical of a given register could be indicative of their language development.

Key words: syntactic functions, formulaic language, lexical bundles, argumentative essay genre, native and non-native speakers of English

1. INTRODUCTION

Researchers and educators in English as a second/foreign language continue to seek new pedagogical tools to help nonnative English speakers develop their academic writing skills in order to become competent members of the international English-medium academic community. With the goal of meeting nonnative academics' particular needs, a growing

number of studies have integrated genre and corpus analytic approaches to the study of academic discourse (L. Flowerdew, 2005). Numerous corpus researchers, for example, have identified formulaic sequences specific to academic genres (e.g., Biber & Barbieri, 2007; Pérez-Llantada, 2014). The formulaic sequences have then been utilized as a means to compare native and nonnative and/or expert and novice writings (e.g., Chen & Baker, 2010; Tribble, 2011). This line of research generally uses published research articles as exemplar texts to compare with various types of academic writing produced by students and/or nonnatives (e.g., Hyland, 2008a; Wei & Lei, 2011). The findings of these studies, while meaningful, remain inconclusive, partly due to their use of texts in different academic genres, which affects the choice and usage of formulaic sequences (Pérez-Llantada, 2014).

A common type of formulaic language is *lexical bundles* (LBs), which are the most frequent recurrent sequences of three or more words in a register or genre; for example, *the end of the* and *in the case of* are common lexical bundles in academic genres (Biber, Johansson, Leech, Conrad, & Finegan, 1999). The previous studies have considered phrases and clauses as the main structures of LBs; however, these boundaries might not always be clear because a bundle's last word is often the first word of another structure. In other words, such fragmented sequences do not stand alone, but are embedded in larger structures within a sentence. The scope of structural LB investigation could be extended, therefore, to the sentence level, enabling researchers to examine different syntactic roles of bundles within a sentence (Cortes, 2015). What's interesting is that, because LBs are determined solely on the basis of frequency, they are generally fragmented phrases or clauses that are embedded in other structures, and consequently, even the same LB can appear in different syntactic roles. Such investigation would provide a more detailed picture of how L1- and L2-English writers use bundles in a register.

In addition, the present study focuses on one specific academic genre, the argumentative essay. Such essays can employ linguistic and conceptual elements that are characteristic of a range of genres. For instance, they may include nominalizations, longer argumentative structures, and critical analysis, which are typical of academic writing, but they can also appropriately utilize conventions of spoken language such as short sentences, more paratactic and fewer hypotactic sentences, and personal statements (Jaworska, Krummes, & Ensslin, 2015). And although it is the genre most frequently required of writers at the university level and in EFL contexts (e.g., Mei, 2006; Wingate, 2012), little research has investigated how L1 and L2 developing writers use LBs in their argumentative writing. This study compares the use of LBs in argumentative essays written on the same topics by native and nonnative writers to provide a detailed picture of how each group deploys LBs in this genre.

2. REVIEW OF THE LITERATURE

Corpus-based and genre-based analyses have been used to complement each other in a number of recent studies (e.g., Belcher, 2012; Charles, 2007; L. Flowerdew, 2005). For example, corpus-based methodologies that are informed by a genre-based approach can deal with larger textual units, addressing an important criticism of corpus analysis, which is that its focus is limited to the level of segmented concordance lines or sentences (e.g., L. Flowerdew, 2005, 2016; J. Flowerdew & R. Forest, 2009). At the same time, genre theory has also benefited from corpus linguistics approaches that can draw on a large number of texts from corpora to better understand genres (Kandil & Belcher, 2011). A recent focus of interest that draws on genre- and corpus-based approaches is the use of formulaic language in specific genres. Corpus linguistics research has explored systematic variations in formulaic sequence use across different genres (e.g., Biber & Barbieri, 2007; Hyland, 2008a, 2012; Pérez-Llantada, 2014; Salazar, 2014; Shin, Cortes, & Yoo, 2018; Shin & Kim 2017). The specific multiword sequences of formulaic language vary according to context but, overall, serve not only to contribute to the meaning and coherence of texts (e.g., Hyland, 2012; Li & Schmitt, 2009), but to help writers perform as competent members of a target discourse community (e.g., Wray, 2002). Scholars have claimed that knowledge of genre-specific sequences contributes to communicative efficiency and fluency in language processing and production (e.g., Arnon & Snider, 2010; Ellis, Simpson-Vlach, & Maynard, 2008; Schmitt, 2004; Tremblay, Derwing, Libben, & Westbury, 2011; Wray, 2002).

Because formulaic language is necessary to text coherence, some scholars consider it to provide the building blocks of discourse (e.g., Biber & Barbieri, 2007; Hyland, 2008b; Leńko-Szymańska, 2014). A common type of formulaic language is *lexical bundles*, which are the most frequent recurrent sequences of three or more words in a register or genre (Biber et al., 1999). Scholars use different frequency cut-off points to identify lexical bundles; a normal range is 10–40 times per million words for four-word bundles, which must occur across five or more texts to avoid idiosyncrasies (Biber, Conrad, & Cortes, 2004). Consequently, lexical bundles are structurally incomplete. They are, however, semantically transparent, serving important functions in both spoken and written discourse (e.g., Biber & Barbieri, 2007; Biber et al., 1999, 2004, 2011).

The findings of studies by Biber and his colleagues counter the traditional account that academic writing is grammatically complex with elaborated clausal structures, which is based on the analysis of T-units (i.e., number of dependent clauses per total clauses; Beers & Nagy, 2011; Inagaki, & Kim,1988; Kroll, 1977; Wolfe-Quintero). The corpus-based studies by Biber et al. have argued that the T-unit is not the best measure of written complexity, because it assesses the use of clausal subordination, which, they claim, characterizes conversation better than academic writing. They demonstrated that academic

writing is structurally *compressed*, with phrasal modifiers embedded in noun phrases, while face-to-face conversation is more *elaborated* with subordinate clauses (Biber & Gray, 2010, p. 7). The reasons for the greater structural compression of academic writing are not only that it can be planned and edited to a much greater extent than speech, but that it is often required to be both more detailed and more concise (Staples, Egbert, Biber, & Gray, 2016). It is generally agreed that all novice academics (L1 and L2) need to learn academic discourse conventions, which include accepted lexical bundle usage, in order to fully participate in the academic community. What is less clear is how novice academic writers develop into experts. One question is whether the development of the ability to use discourse conventions appropriately progresses along the same path for all writers. In the past few years, an increasing number of researchers have used lexical bundles as a tool to investigate academic language development, and the extent to which L2 learners have difficulty in the use of bundles has generated much discussion.

Many corpus-based studies have compared native and nonnative LB use in academic written contexts (e.g., Ädel & Erman, 2012; Salazar, 2014). Hyland (2008a), for instance, compared LBs in L2 English student writing (theses and dissertations) to published research articles. He found differences that he claimed were due to "genre variations" (p. 50). The differences he observed, however, could be partly due to writer characteristics, but he did not consider writers' language background (i.e., as native or nonnative English speakers) as a factor. On the other hand, Ädel and Erman (2012) demonstrated divergent patterns of LB usage specific to L1 versus L2 undergraduate writers, claiming that such differences were derived from language background. However, the patterns may also have been affected by differences in the corpora used in their study, which included different L1 and L2 writing genres and tasks. Academic writing is greatly influenced by both writing prompts and author profiles, which determine genre and task-type (e.g., Friginal, Li, & Weigle, 2014); however, very few studies have taken such factors into consideration.

There are also studies that report syntactic developmental trends in university level texts written by native English writers. Biber, Gray, and Poonpon (2011) analyzed the syntactic functions of grammatical features, which they ranked in terms of grammatical features' contribution to syntactic complexity, in research articles and conversation produced by native speakers of English. They hypothesized a sequence of developmental stages for L2 writing to predict the order in which L2 learners acquire particular features. Biber et al. speculated that L2 writers will use complexity features common in conversation before they use the complexity features in academic writing. Specifically, they proposed that L2 writers will tend to acquire finite dependent clauses at earlier stages, nonfinite dependent clauses at intermediate stages, and a variety of phrase types within noun phrases at later stages. Although their study did not analyze lexical bundles, they paved the way for the exploratory use of this approach in the production of other linguistic features such as

lexical bundles.

Several researchers have since provided empirical support for Biber et al.'s (2011) hypothesized developmental sequence through studies on advanced academic writing by L2 writers. For example, Parkinson and Musgrave (2014) focused on noun phrase complexity. They examined academic writing produced by graduate L2 writers at two proficiency levels. The less proficient group overused attributive adjectives, which Biber et al. expected to be acquired at an early stage, whereas the more proficient group used noun phrases as postmodifiers, a use that Biber et al. suggested would be acquired at a later stage. Parkinson and Musgrave's findings, however, might not be conclusive because their groups also produced different registers of academic writing. The lower proficiency group's texts were argument essays while the higher proficiency group's texts were discipline-specific writing in the field of TESOL/applied linguistics; each of these registers requires specific structures and writing skills that are developed at different stages (Nesi & Gardner, 2012). Future analyses with two more comparable corpora controlled for register matter are needed to demonstrate whether the differences found in this study are simply because of the learner writers' different levels, which is an issue consistently found in this field, as mentioned above.

Another study, conducted by Staples et al. (2016), is among the very few developmental studies with L1 writers at the university level, from undergraduate to graduate. Staples et al. demonstrated developmental trends in university level texts written by native English writers. Using the British Academic Written English corpus, they found that student writing tends to use more phrasal complexity but less clausal complexity (especially finite dependent clauses) as the writers' academic level increases. Based on their findings, Staples et al. argued that university-level student writers are still developing their ability to use grammatical structures as they become familiar with and able to use the discourse styles of academia, and that this type of development is distinct from their learning of new genres and of disciplinary expectations. They concluded, as well, that L1 writers' development during university education shares some common features with L2 writers' development in the same context.

The present study then extends the internal structural analysis of LBs in the literature by focusing on the syntactic roles of structures in which LBs are constructed. The study uses two parallel corpora, one of native English-speaking students' writing samples and one of native Korean-speaking English-as-a-foreign-language (EFL) students' writing samples, controlled for register (i.e., argumentative essays) and writing prompt (i.e., same topics and time constraints). To this end, the following research question is posed:

> How do native and nonnative English-speaking first-year university writers use lexical bundles in their writing in terms of syntactic functions in response

to the same writing topics?

3. METHODOLOGY

3.1. Corpus Data

This study used native and nonnative English corpora of argumentative essays written in response to the same writing prompts by university freshmen at the very beginning of their first semester. The learner corpus (LC, hereafter) was built on English writing samples from entering freshmen at a university in Korea. A total of 1,408 students wrote argumentative essays as part of the placement test for mandatory first-year English courses, amounting to 491,800 words. They were instructed to write an essay on a given topic in paragraph form in 50 minutes. One of eight topics was given to each student. One example topic is:

> It has been said, "Not everything that is learned is contained in books." Compare and contrast knowledge gained from experience with knowledge gained from books. Which source is more important? Use specific reasons and examples to support your answer.

The native corpus (NC, hereafter) was built on writing samples from L1 English first-year students at a large public university in the southeastern United States. In the first week of freshman composition courses in 2017, the students were asked to write essays as a diagnostic test. The writing prompts and time constraints were the same as those for the essays in the nonnative corpus. To ascertain the students' first language, they were asked to provide some demographic information, and essays written by students with L1s other than English were excluded from the corpus.

TABLE 1

Description of the Two Corpora

Corpora	Number of Essays	Mean Length of Essays (Words)	Total Corpus Size (Words)
Learner (LC)	1,408	349.3	491,800
Native (NC)	1,414	346.9	490,610

The native corpus contains 1,414 essays, of 346.9 words on average, amounting to 490,610 words, with a very similar average essay length and total number of words to that of the learner corpus, as in Table 1.

Considering that the two corpora each contain approximately 490,000 words, raw frequencies were used without converting them to a normalized rate. The frequencies of all the bundle types in the two corpora were tested for statistical significance using log-likelihood tests.[1]

3.2. The Construction of Lexical Bundles in Context

First, 4-word LBs were identified in the native and nonnative corpora, using the commercial concordance software *AntConc* (Anthony, 2014). The study analyzes 4-word sequences "because they are far more common than 5-word strings and offer a clearer range of structures and functions than 3-word bundles" (Hyland, 2008b, p. 8). Following Biber et al. (1999), the frequency threshold was set at 10 times in the native and nonnative corpora, which contains approximately half a million words each, and the range threshold at a minimum of five different texts.

The LBs identified in each corpus were then categorized using structural and functional taxonomies for the classification of LBs (Biber et al., 1999, 2004). The structural categorization involved identifying types of internal structural units: clausal (i.e., VP-based bundles) and phrasal (i.e., NP- and PP-based bundles). VP-based bundles include word sequences with a verb component. NP-based bundles refer to those including nominal phrases with *of*-phrase fragments and post-modifier fragments, and PP-based bundles comprise a preposition followed by an NP fragment. In order to investigate how native and nonnative writers construct LBs in context, all the bundles identified in each corpus were analyzed in terms of their syntactic roles (Cortes, 2015). With respect to verb-phrase LBs (e.g., *it is important to*), they were first structurally categorized depending on whether they occur in a main verb phrase or a dependent clause. Those in the dependent clause category were then subcategorized by the syntactic roles played by the clause; for example, adverbial (e.g., *Although it is important to*), complement (e.g., *It is known that it is important to*), and noun modifier (e.g., *This is the reason why it is important to*), or any other syntactic role emerging in context.

4. RESULTS AND DISCUSSION

This section presents the lexical bundles identified in the argumentative essays produced by the native and nonnative student writers. Appendix A (Shin, 2018) provides the final list of bundles from both corpora (NC: 146 types, LC: 156 types) after topic-dependent

[1] I used Paul Rayson's log-likelihood calculator from http://ucrel.lancs.ac.uk/llwizard.html.

bundles that directly quoted the given essay prompts were removed. The LB used by each group were first categorized according to three main structural types: NP-based, PP-based, and VP-based bundles. In what follows, each LB structural type (i.e., VP-, NP-, and PP-based) is presented in terms of its subcategories in both corpora.

4.1. VP-based Bundles

VP-based bundles comprise the largest proportion of the total number of bundles identified in both corpora (NC: 65.7%, LC: 69.2%). VP-based bundles were first categorized according to whether they are embedded in the main clause or a dependent clause. Those in the latter group (NC: 718 tokens, LC: 813 tokens) were then subcategorized by the syntactic roles served by the clause.

Table 2 shows the roles of VP-based bundles (i.e., those occurring in dependent clauses) with the results of log-likelihood tests comparing the numbers of tokens for each role in the two corpora. The tests showed significant differences in eight syntactic roles served by VP-based bundles in both corpora. The learners used seven syntactic roles more frequently, including finite complement clauses controlled by a common V (extremely common verbs in conversation such as *think*, *know*, and *say*; Biber et al., 2011, p. 30), and by a copula (particularly *be*-verb), and WH relative clauses.

TABLE 2

Distribution of Syntactic Roles of VP-based Bundles in NC and LC

Syntactic Roles	NC	LC
Finite complement clause (CC) controlled by common V****	25(3.5%)	75(9.2%)
Finite CC controlled by V****	115(16%)	41(5%)
Finite CC controlled by copula***	9(1.2%)	30(3.7%)
Finite CC controlled by predicative Adj	5(0.7%)	12(1.5%)
Finite CC controlled by N	5(0.7%)	1(0.1%)
who relative clause****	5(0.7%)	27(3.3%)
which relative clause****	10(1.4%)	38(4.7%)
that relative clause	79(11%)	65(8%)
Finite adverbial clause	322(44.8%)	373(45.9%)
Nonfinite CC controlled by Common V	10(1.4%)	18(2.2%)
Nonfinite CC controlled by V	25(3.5%)	19(2.3%)
Nonfinite CC controlled by copula	2(0.3%)	1(0.1%)
Nonfinite CC controlled by predicative Adj	15(2.1%)	11(1.3%)
Nonfinite CC controlled by N	1(0.1%)	-
Nonfinite relative clause*	5(0.7%)	15(1.8%)
Nonfinite adverbial clause	44(6.1%)	33(4%)
Comparative clause**	17(2.4%)	37(4.5%)
Other	24(3.3%)	17(2.1%)
Total	718(100%)	813(100%)

Note. CC: complement clause, V: verb, Adj: adjective, N: noun; * = significant at $p < .05$; ** = significant at $p < .01$; *** = significant at $p < .001$; **** = significant at $p < .0001$.

While both student writer groups use VP bundles in several types of syntactic roles, both groups show excessive use of them in certain roles; one that stands out is the finite adverbial clause. In both corpora, approximately 45% of all VP bundles fills this particular syntactic role (NC: 44.8%, LC: 45.9%). However, the types of these clauses favored by each group are not necessarily the same. Table 3 lists the subordinators in adverbial clauses found in NC and LC. As the table shows, the top three are identical: *when* is the most frequent in both corpora, and comprises almost half of the total tokens of subordinators in adverbial clauses in NC. The second top subordinator is *if*, which is used to more or less the same extent by the two groups at about 25%, followed by *because*, which is particularly used more often by the learners.

TABLE 3

Types of Finite Adverbial Clauses in NC and LC

Subordinators	NC	LC
when	156(48.4%)	119(31.9%)
if	83(25.8%)	95(25.5%)
because	40(12.4%)	74(19.8%)
although	20(6.2%)	12(3.2%)
while	16(5%)	4(1.1%)
though	3(0.9%)	2(0.5%)
even though	2(0.6%)	12(3.2%)
even if	2(0.6%)	-
as	-	34(9.1%)
since	-	19(4.8%)
whenever	-	3(0.8%)
Total	322(100%)	373(100%)

It should be noted that several LBs themselves contain *when*-, *if*-, or *because*-clause fragments, which constitute a large portion of the total tokens reported in Table 3. In the native corpus, 62.1% (200 tokens) are such bundles: three LBs with embedded *when* fragments (e.g., *when it comes to*, *when I was in*) amount to 141 tokens, and four LBs with *if* (e.g., *if you do not*, *if I had to*) account for 59 tokens. Similarly, in the learner corpus, LBs with adverbial clause fragments comprise 60% (224 tokens) of these bundles, with four different subordinators: *when* (119 tokens), *if* (55 tokens), *because* (25 tokens), and *as* (25 tokens).

The following examples, taken from essays on the same topic in the two corpora, demonstrate this use of LBs (indicated in bold), with *although* in NC (1) and with *since* in LC (2).

(1) Although Gwinnett **is one of the** most populated counties in Georgia, it is vastly spread out and this problem makes it near impossible to get from

one section to the other without an automobile. (NC, topic 1)

(2) This aspect may bring a huge loss since Pusan **is one of the** major tourist sites for foreigners. (LC, topic 1)

While the top three adverbial clauses are of the same type in both corpora, the rest show features specific to each group, and the learners employ more varied types of clauses in which LBs are embedded (NC: 5 types, LC: 7 types). As shown in Table 3 above, the remaining adverbial clauses in NC are restricted to concessive clauses with *although*, *while*, *though*, *even though*, and *even if*. Compared to the natives, the learners use only a small number of concessive clauses, and especially few with *while*, although they use *even though* more frequently than do the natives. The learners instead tend to use causal clauses including *as* and *since*, neither of which were found in NC.

Overall, both groups predominantly employ the syntactic roles of finite dependent clauses (NC: 80%, LC: 81.4%, of all VP bundles). Recent corpus-based studies have demonstrated that academic writing is structurally "compressed," with complex noun phrase constituents and phrases, while face-to-face conversation is more "elaborated" with subordinate clauses (Biber & Gray, 2010, p. 7). In particular, adverbial clauses are the most common feature of interpersonal spoken registers (e.g., Biber et al., 1999, 2011). The fact that finite adverbial clauses comprise the largest type found in both corpora appears to reflect the writers' status as novice academic writers (native and nonnative alike) who are just transitioning to the university level. Additionally, in some cases, learners erroneously use adverbial clauses in ways indicative of learner language, as in (3–4), where the dependent clauses stand alone, fragmented, without a connection to the main clause (LBs in bold).

(3) Because degree **is one of the** objective reason to hire job. (LC, topic 2)

(4) Even though **it doesn't mean that** current older people are less intelligent than past older people. (LC, topic 8)

Such errors are mostly limited to *because* clauses; few were found with other subordinators. This finding corresponds to one reported by Yoon and Yoo (2011), whose participants, Korean learners of English, frequently produced such fragmentary sentences. The authors argued that L1-Korean students tend to regard *because* as a conjunctive adverb, possibly due to negative transfer from their first language; the word corresponding to *because* in Korean is used as an adverb.

While both groups use finite complement clauses at similar rates, nonfinite complement clauses, mostly realized in the form of *to*-clauses, are used relatively more by the native writers (NC: 14.2%, LC: 11.9%). This type of structure is generally less frequent in spoken

registers (Biber et al., 1999) unless it occurs in the combination of *want* + *to* clause, which is extremely common in conversation (p. 711). A close examination shows that *want* + *to* clauses are prevalent in both corpora. Moreover, some of the LBs identified in both corpora themselves include *want* + *to* (4 LBs in NC, 10 in LC), with three shared bundles (i.e., *want to be a*, *do not want to*, and *if you want to*). The following examples illustrate the use of the shared bundle *to go to the* combined with *want* in NC (5) and in LC (6).

(5) This may seem like a good thing because every parent wants their child **to go to the** best school possible and become as successful as possible. (NC, topic 1)
(6) This is the reason why our parents want us **to go to the** university. (LC, topic 8)

The next most frequent role in both corpora is that of the finite complement clause (CC) preceded by a verb. The types of verbs used by each group, however, differ significantly ($p < .0001$). The learners are most likely to combine the finite CC with a common V such as *think*, *say*, or *know* (9.2%), all frequently used in conversation (Biber et al., 2011). Examples (7) and (8) show a finite CC controlled by V (*believe*) from the native corpus and a finite CC controlled by common V (*think*) from the learner corpus, respectively.

(7) I know that there's a really slim chance of college ever being free, but a huge step would be to at least make it affordable. I believe the cost of college **is one of the** main reasons students don't attend. (NC, topic 2)
(8) Chance of education. I think that **is one of the** most powerful point of big city. (LC, topic 2)

It should be emphasized that the grammatical complexity of the VP bundles that occur in dependent clauses is not always the same. Rather, their complexity differs according to the adjoining elements of the bundles. That is, while elaborated dependent clauses are typical of spoken language, those serving as constituents in an NP are strongly favored in academic writing (except for *that*-relative clauses; Biber et al., 1999, 2011). For example, adverbial subordination and complements controlled by verbs are associated with clausal syntax, while complements controlled by nouns are associated with phrasal syntax. As in Table 2 above, LBs serve three syntactic roles that are constituents in NPs: (1) finite CC controlled by N; (2) nonfinite CC controlled by N; (3) relative clause (WH relative clause). There was no significant difference in the very low frequency with which the two groups use LBs in the first two roles (NC: 0.8%, LC: 0.1%). Although the difference is very small, the native writers use more noun complement clauses than the learners. The examples

below demonstrate the use of nonfinite LBs in a complement clause controlled by a noun, *decision*, in NC (9), and in a relative clause postmodifying a noun, *way*, in LC (10).

(9) When I was in the 6th grade, I made the decision **to go to the** store with my friends. (NC, topic 5)
(10) Edison had made the world brighter than ever been before by inventing light bulb. By breaking thousands of prototypes, he eventually got to the way **to solve the problem** he was facing. (LC, topic 5)

The other syntactic role of LBs related to phrasal syntax is that of the WH relative clause. As shown in Table 2, the learners, surprisingly, use more LBs in these structures than the natives (NC: 2.1%, LC: 8%). Biber (1995), who conducted multidimensional analyses for different languages including English and Korean, demonstrated that nominal modifiers such as relative clauses (except for *that* relative clauses) are generally typical of academic writing across languages. The following examples show the use of bundles in WH relative clauses in NC (11) and LC (12).

(11) Writing can be an extremely powerful way to be heard by groups who you may **not be able to** reach otherwise. (NC, topic 3)
(12) However, public education can not meet the needs of parents and **students who want to** enter high class universities anymore. (LC, topic 1)

Although previous corpus-based studies have argued that WH-relative clauses as post nominal modifiers are strongly associated with academic writing, this study's close examination of the use of the LBs in the two corpora shows that student writers' usages do not necessarily conform to the norms of academic prose. This is especially the case for the learner corpus, which shows the frequent use of LBs in *who*-clauses (over 40% of WH-clauses). According to Biber et al. (1999, pp. 609–611), *who*-clauses and *which*-clauses have notably different distributions across registers in that the former occur only with animate (usually personal/human) head nouns and the latter usually occur with inanimate head nouns. Thus, a communicative focus on humans, as in the registers of news and fiction, results in the frequent use of relative clauses with the pronoun *who*. In contrast, *which*-clauses are the most frequently used WH-clause type in academic prose, where animate references tend to be few. The learners' frequent use of *who*-clauses therefore appears to be a deviation from typical academic prose.

4.2. NP-based Bundles

The native and nonnative writers show distinctive differences in the use of NP-based bundles (NC: 14.4%, LC: 15.4%). Table 4 presents the subcategories of syntactic roles with the results of the log-likelihood tests for each role.

TABLE 4

Distribution of Syntactic Roles of NP-based LBs in NC and LC

Syntactic Role	NC	LC
Subject****	96(21.7%)	234(41.9%)
Subject predicative****	77(17.4%)	160(28.7%)
Direct object	96(21.7%)	85(15.2%)
Indirect object	3(0.7%)	4(0.7%)
Object predicative	1(0.2%)	3(0.5%)
Agent in passive voice	0(0%)	2(0.3%)
PP as adverbial****	120(27.1%)	58(10.4%)
of-phrase as postmodifier****	49(11.1%)	5(0.9%)
Relative clause	0(0%)	2(0.3%)
Other	1(0.2%)	5(0.9%)
Total	443(100%)	558(100%)

Note. **** = significant at $p < .0001$.

As shown in Table 4, significant differences were found in the frequency of the four syntactic roles played by NP-based bundles in both corpora. First, the learners mostly use NP bundles as subjects or subject predicatives, which together account for about 69.1% (subject: 41.9%, subject predicative: 27.2%). The study of Cortes (2015), who examined syntactic roles of LBs in published research articles, found that NP bundles functioned as direct objects most frequently (36%), followed by subjects (20%), subject predicatives (9%), and adverbials (6%). One similarity between the professional writing corpus in her study and the NC corpus in this study, despite the different academic genres, is that approximately 20% of the NP bundles function as subjects in both. In contrast, approximately 40% of the NP bundles in this study's LC take the role of subject. The following examples illustrate the syntactic roles of NP bundles in the two corpora. LBs play the role of object in NC (13), and subject and subject predicative in LC (14–15). Note the use of *because* with a comma in (14), which lends further support to the argument that L1-Korean students consider English *because* an adverb (Yoon & Yoo, 2011).

(13) Alex Rodriguez was a star baseball player who broke out into the scene and shocked **a lot of people** because of his talent. (NC, topic 5)

(14) Because, **a lot of people** use public transportation with many reasons. (LC, topic 1)

(15) There are **a lot of people** who don't know what they want to do until they become a grownup. (LC, topic 2)

The high proportion of NP bundles as subject predicatives in LC might be related to the abundance of the copula *be*-verb in the learners' writing. Chen and Baker (2016), who examined LBs in L2 academic writing, found that one-third of the LBs included *be*-verbs. The authors argued that the overuse of *be*-verbs came from the learners' heavy reliance on existential *there*-constructions (e.g., *there are so many*), and that both tendencies made their writing style "simplistic and verbose" (p. 866). In the same vein, the LC in the current study includes a considerable number of LBs with embedded *be*-verbs (36 types, 23.1%) and/or *there*-constructions (17 types, 10.9%). Moreover, the analysis of co-structures of LBs shows that the tendency is not limited to LBs, but extends to LB co-structures as well, as in examples above.

A related account for the frequent use of LBs as subjects points to the learners' use of sentence-initial bundles. Recent studies (Li, Franken, & Wu, in press) compared the position of bundles in native and nonnative postgraduate academic writing, demonstrating that nonnative writers were most likely to start sentences with LBs. Follow-up interviews conducted by Li et al. (2018) with five L1-Chinese postgraduate writers suggested possible reasons. Such reasons include previous learning experience (e.g., teachers' overemphasis on English conjunctions and formulaic sequences as a strategy for cohesion and coherence in academic writing), noticing in reading (expressions frequently occurring in the initial place of sentences in their course books), and a lack of rhetorical confidence – they may simply be more comfortable starting with familiar expressions to minimize the risk of making mistakes. Lending further support to this argument, the learners' frequent use (over 40% of the NP bundles) of LBs as subjects found in the present study seems to be related to the reasons suggested by Li et al.

In comparison, the natives employ NP-based bundles within PPs functioning as adverbials or postnominal modifiers, which together comprise 38.1% of their NP-based bundle uses. In particular, the native writers often embed NP bundles in *of*-phrases (49 tokens, 11.1%) in compressed structures typical of academic prose while the learners rarely do so (5 tokens, 0.9%). Examples of LBs in PPs functioning as adverbials in NC appear in (16) and in LC in (17).

(16) There are exceptions to every blanket statement; however, mistakes are endlessly inevitable for **the majority of the** human race and tend to follow people for their entire lives despite the attention to the consequences of certain actions. (NC, topic 5)

(17) Some may have nothing on their mind at **the end of the** class. For

example, many Korean high school students are having a hard time concentrating in classes because they have to follow traditional ways of school teaching. (LC, topic 7)

Examples (18) and (19) demonstrate NP bundles embedded in PPs serving as postnominal modifiers in NC and LC, respectively. Note that two bundles are used in a row in (18); the first (underlined) is an NP with an *of*-phrase fragment and the second (bold) is embedded in the *of*-phrase.

(18) They both allow for a better understanding of **the world around us**. (NC, topic 3)
(19) One of **the most important thing** about city is how people in town are able to travel out to other place easily. (LC, topic 1)

4.3. PP-based Bundles

As for PP-based bundles (NC: 18.5%, LC: 14.7%), both native and nonnative writers mainly use these LBs as adverbials, as shown in Table 5. In some cases, the natives also use PP bundles as post nominal modifiers (4.6%), which the learners rarely do (0.9%).

TABLE 5
Distribution of Syntactic Roles of PP-based Bundles in NC and in LC

Syntactic Role	NC	LC
Adverbial	576(95.4%)	447(99.1%)
Post nominal modifier	28(4.6%)	4(0.9%)
Total	604(100%)	451(100%)

As in Table 5, the writers do use a small number of PP bundles as post modifiers, but they are rare in the NC and even rarer in the LC. Examples (20–21) illustrate the use of the same PP bundle, *all over the world*, serving as a post nominal modifier in responses to the same topic in NC (20) and LC (21).

(20) Clarkston is known as a refugee hub with thousands of different languages being spoken and people from **all over the world** entering the United States unable to speak English. (NC, topic 1)
(21) Nowadays, Incheon is becoming a global city by holding several international events, also lots of countries **all over the world** started to focusing on development of Incheon. (LC, topic 1)

Another PP bundle, *in the middle of*, is frequently used as a postnominal modifier in NC as shown in (22), but always used as an adverbial in LC as in (23).

(22) It is a large cluster of rocks right **in the middle of** the Chattahoochee that creates a natural waterslide. (NC, topic 6)
(23) For example, every month 2th and 7th, consumer direct market is held **in the middle of** the city. (LC, topic 6)

Previous studies have shown that academic writing relies on phrasal bundles, many of which in fact occur in multiple PPs as postnominal modifiers embedded in an NP such as *the presence of layered structures at the borderline of cell territories* (multiple PPs underlined; Biber et al., 2011, p. 31). Several scholars have suggested that such styles of discourse, which only occur in certain circumstances of formal writing, are not naturally acquired, and even native writers may rarely (or never) use them, especially before adulthood (Biber et al., 2011; Staples et al., 2016). Instead, novice academic writers gradually learn to produce such complex structures over the course of their university education (Staples et al., 2016). Given that the student writers in this study were entering undergraduates, it can be assumed that they had yet to receive formal writing instruction at the university level. It therefore seems reasonable that the student writers, and particularly the learners, seldom produce multiple PP bundles as postnominal modifiers. Furthermore, it should be noted that the types of PP bundles identified in the two corpora deviate from the types generally found in academic prose. Many of the LBs in both corpora (e.g., *with that being said*, *on a daily basis*, and *all over the world*) are more typical of spoken language, which may be related to their use as adverbials rather than post nominal modifiers.

5. CONCLUSION

The present study explored the use of lexical bundles in association with the syntactic roles they play in a clause. All the bundles identified in both corpora were categorized in terms of syntactic roles. Overall, the results show that native and nonnative incoming college students display generally similar patterns of using lexical bundles in certain syntactic roles, but at the same time, their patterns do diverge to some extent, with some uses unique to or more common to each group. The similarities are presumably due to their status as novice academic writers. In common, both groups predominantly use VP-based bundles in elaborated clausal structures, especially in two specific structures: finite adverbial clauses and finite complement clauses controlled by verbs, which together

account for more than 60% of the syntactic structure types in each corpus (NC: 64.3%, LC: 60.1%). These grammatical features are the most common in interpersonal spoken registers (e.g., Biber et al., 1999, 2011).

As for NP-based bundles, the two language groups employed them in divergent syntactic roles. One difference was the learners' excessive use of NP bundles as subjects or subject predicatives (about 70% together) and far less frequent use of them as objects, compared to their native counterparts. This pattern appears to be related to the learners' frequent use of copula *be*-verbs (Chen & Baker, 2016) and/or their tendency to place formulaic language in the initial position of sentences (Li et al., 2018). Another difference involves the native writers' use of NP bundles embedded in *of*-phrases functioning as postmodifiers, which is a pattern closer to the norms of expert academic prose. The learners, however, rarely used NP-based bundles in this way in their writing. In addition, the learners almost always used PP-based LBs functioning as adverbials whereas the natives, in some cases, produced compressed structures embedding LBs in multiple PPs as post nominal modifiers. The analysis of the internal structures of LBs showed similar uses of phrasal LBs in both corpora, but that of the syntactic roles of LBs along with co-occurring structures revealed features unique to each group.

As mentioned earlier, Biber et al. (2011) ranked grammatical features by their contribution to linguistic complexity, using written and spoken native speaker data (research articles and face-to-face conversation). Based on their findings, Biber et al. proposed a developmental progression in which L2 academic writers produce clausal complexity before they use the complexity features common in academic writing. Specifically, they argued that L2 writers tend to acquire finite dependent clauses at earlier stages, followed by nonfinite dependent clauses and a variety of phrase types at later stages. In this developmental sequence, both the native and nonnative student writers in the current study, who predominantly use finite dependent clauses, appear to be at the initial stage. It is important, however, to bear in mind that many previous studies that document the grammatical features of academic writing, including Biber et al.'s (2011), used published research articles as their academic writing data. The characteristics of this formal written genre may not apply to other types of academic writing.

Note that preferences for complexity features differ *within* academic writing, as they are shaped by communicative expectations in a given genre and discipline (e.g., Hyland, 2012; Staples et al., 2016). In particular, argumentative essays "are normally not recognized as fully fledged academic texts, since they lack references or a rigid mesostructure" (Jaworska et al., 2015, p. 508). In this regard, the clausal features manifested in both corpora should be, to some extent, interpreted as features of this particular genre, in addition to being features of novice academic writing. Among the handful of studies on LBs in argumentative essays, most, surprisingly, do not take such factors into consideration,

instead considering the prevalence of clausal complexity in this genre a feature of L2 writing (e.g., Bychkovska & Lee, 2017; Staples et al., 2013). Future research on this topic should include argumentative essays by different groups of writers to draw a fuller picture of LB use in this register by examining to what extent clausal complexity is typical of it.

Furthermore, the findings have pedagogical implications. To develop their academic writing skills, both native and nonnative English-speaking novice student writers must learn how to use multiword sequences appropriately in context. They can therefore benefit from explicit instruction on specific lexical bundles and the structures in which they are embedded and with which they occur. In addition, the finding that the nonnative writers used more academic-register features than the native writers (e.g., post nominal modifiers) as well as more spoken features (e.g., colloquial expressions) indicates that explicit instruction on features specific to academic writing would be beneficial for English learners in particular.

As discussed, there exists relatively little research using parallel corpora of native and nonnative academic genres, other than published research articles. The study adds to our understanding of the use of lexical bundles by different language groups and provides useful information for teaching academic writing to novice academic writers. Future research targeting the same type of academic writing produced by different language groups would provide a more concrete picture of the groups' formulaic language use, as well as of the features specific to the register.

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*, 81–92.

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer software]. Tokyo, Japan: Waseda University [Available from http://www.laurenceanthony.net/].

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*(1), 67-82.

Belcher, D. (2012). Considering what we know and need to know about second language writing. *Applied Linguistics Review, 3*(1), 131-150.

Beers, S., & Nagy, W. (2011). Writing development in four genres from grades three to seven: Syntactic complexity and genre differentiation. *Reading and Writing, 24*, 183-202.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge, England: Cambridge University Press.

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

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

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes, 9*, 2-20.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristic of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5-35.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow, UK: 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. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology, 14*(2), 30–49.

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. (2015). *Analyzing the syntactic roles and semantic prosodies and preferences of non-trigger lexical bundles*. Paper presented at PRISEAL 3 (Publishing and Presenting Internationally: Issues for Speakers of English as an Additional Language), Coimbra, Portugal.

Ellis, N., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 43*(3), 375-396.

Flowerdew, J., & Forest, R. (2009). Schematic structure and lexico-grammatical realization in corpus-based genre analysis. In M. Charles, D. Pecorari & S. Hunston (Eds.), *Academic writing: At the interface of corpus and discourse* (pp. 15-36). London: Continuum.

Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for Specific Purposes, 24*, 321-332.

Flowerdew, L. (2016). A genre-inspired and lexio-grammatical approach for helping postgraduate students craft research grant proposals. *English for Specific Purposes, 42*, 1-12.

Friginal, E., Li, M., & Weigle, S. (2014). Revisiting multiple profiles of leaner compositions: A comparison of highly rated NS and NNS essays. *Journal of*

Second Language Writing, 23, 1-16.

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

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

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics, 32*, 150-169.

Jaworska, S., Krummes, C., & Ensslin, A. (2015). Formulaic sequences in native and non-native argumentative writing in German. *International Journal of Corpus Linguistics, 20*(4), 500-525.

Kandil, M., & Belcher, D. (2011). ESP and corpus-informed critical discourse analysis: Understanding the power of genres of power. In D. Belcher, A. Johns & B. Paltridge (Eds.), *New directions in English for specific purposes research* (pp. 252-270). Ann Arbor, MI: University of Michigan Press.

Kroll, B. (1977). Ways communicators encode propositions in spoken and written English: A look at subordination and coordination. In E. O. Keenan & T. Bennett (Eds.), *Discourse across time and space (SCOPIL no. 5)* (pp. 69-108). Los Angeles: University of Southern California.

Leńko-Szymańska, A. (2014). The acquisition of formulaic language by EFL learners: A cross-sectional and cross-linguistic perspective. *International Journal of Corpus Linguistics, 19*(2), 225-251.

Li, L., Franken, M., & Wu, S. (in press). Chinese postgraduates explanation of the sources of sentence initial bundles in their thesis writing. *Regional Language Centre Journal.*

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing, 18*, 85-102.

Mei, W. (2006). Creating a contrastive rhetorical stance: Investigating the strategy of problematization in students' argumentation. *RELC Journal, 37*(3), 329-353.

Nesi, H. & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge: Cambridge University Press.

Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for academic purposes students. *Journal of English for Academic Purposes, 14*, 48-59.

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*. Philadelphia, PA: John Benjamins.

Schmitt, N. (Ed.). (2004). *Formulaic sequences: Acquisition, processing, and use*. Amsterdam: John Benjamins.

Shin, Y. (2018). *Lexical bundles in argumentative essays by native and nonnative English-speaking novice academic writers*. Unpublished doctoral dissertation, Georgia State University, Atlanta, GA.

Shin, Y., Cortes, V., and Yoo, I. (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, Y., & Kim, Y. (2017). Using lexical bundles to teach articles to L2 English learners of different proficiencies. *System, 69,* 79-91.

Stapes, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication, 33*(2), 149-183.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning, 61*(2), 569-613.

Tribble, C. (2011). Revisiting apprentice texts: Using lexical bundles to investigate expert and apprentice performance in academic writing. In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 85-108). Amsterdam: John Benjamins.

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

Wingate, U. (2012). 'Argument!' helping students understand what essay writing is about. *Journal of English for Academic Purposes, 11*, 145-154.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1988). *Second language development in writing: Measures of fluency, accuracy, and complexity* (Technical Report No. 17). Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawaii.

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

Yoon, J., & Yoo, I. (2011). An error analysis of English conjunctive adjuncts in Korean college students' writing. *English Teaching, 66*(1), 224-244.

APPENDIX A

Distribution of LBs in NC and LC (Shin, 2018)

Native corpus (146 types, 2783 tokens)				Learner corpus (156 types, 3434 tokens)			
on the other hand	110	**if you want to**	14	**is one of the**	110	**in my high school**	15
when it comes to	109	in the city of	14	there are lots of	105	a high school student	15
disagree with the statement	63	in the heart of	14	**on the other hand**	88	you don't have to	14
to be able to	60	is more likely to	14	**the most important thing**	88	as I mentioned above	14
it would be the	58	it is up to	14	**there are so many**	88	for example there is	14
is one of the	52	the majority of the	14	**I would like to**	78	from now on I	14
is more important than	45	the only way to	14	but I think it is	78	is very hard to	14
the best way to	45	there is no way	14	there are a lot of	62	it is difficult to	14
due to the fact	43	**this is not the**	14	a person I know	54	there are things that	14
the end of the	39	are going to be	13	**is more important than**	53	because they are not	13
a lot of people	38	**but it is not**	13	so I want to	52	but I think the	13
most of the time	36	do not know how	13	**agree with the statement**	51	I am going to	13
the rest of the	33	**in the middle of**	13	for these reasons I	48	in front of the	13
in my opinion I	32	that need to be	13	**a lot of people**	46	it is impossible to	13
agree with the statement	31	the fact that it	13	**a lot of things**	46	my point of view	13
is a lot of	30	**there are a few**	13	**there are many things**	43	the reason why I	13
in a way that	29	**all over the world**	12	**when I was in**	42	there are not enough	13
not be able to	28	do not have the	12	**has a lot of**	40	there are three reasons	13
will be able to	28	**for the first time**	12	**it is hard to**	38	there is a saying	13
does not mean that	27	**has a lot of**	12	**to go to the**	37	which is located in	13
are more likely to	26	how to deal with	12	**disagree with the statement**	36	a person who is	12
for the most part	26	how to do something	12	what I want to	36	and the other is	12
at the same time	25	I do believe that	12	**will be able to**	35	as a result I	12
one of the most	24	if I were to	12	when I was young	34	as you can see	12
there are many things	24	is not always the case	12	**it is true that**	32	at that time I	12
I would like to	23	is the amount of	12	don't know how to	32	because there is no	12
do not agree with	23	**it is easy to**	12	**at the same time**	31	however I think that	12
over and over again	23	on a daily basis	12	**when it comes to**	31	is very famous for	12

APPENDIX A

Continued (Shin, 2018)

Native corpus (146 types, 2783 tokens)				Learner corpus (156 types, 3434 tokens)			
what not to do	23	this is why I	12	have a chance to	30	it is not true	12
there are so many	22	what is going on	12	place to live in	29	most of the people	12
when I was in	22	teach an old dog new tricks	11	**if you want to**	29	one of my friends	12
I was able to	21	**for a long time**	11	there are not many	29	one thing that I	12
I was born in	21	have been able to	11	**it is easy to**	26	the people who are	12
would have to be	21	I do not think	11	therefore if I could	26	thank you for reading	12
I feel as if	20	I would want to	11	**but it is not**	25	**there are a few**	12
in the long run	20	in and out of	11	**one of the most**	25	this is because the	12
with that being said	20	is a great place	11	**for a long time**	24	when I go to	12
to go to the	20	is a part of	11	he or she would	24	will be helpful to	12
a better understanding of	19	is the key to	11	**I was born in**	23	**it doesn't mean that**	11
a lot of things	19	it depends on the	11	to live in my	23	are not good at	11
at a young age	19	**it is hard to**	11	to solve this problem	23	as a result the	11
I am able to	19	**one of the biggest**	11	**want to be a**	23	as soon as possible	11
I believe that the	19	tend to be more	11	why I want to	23	as time goes by	11
it is important to	19	**the most important thing**	11	I strongly believe that	22	because of lack of	11
would be able to	19	the only thing that	11	so if I can	22	have a right to	11
don't get me wrong	18	**there are many ways**	11	the center of the	22	however I believe that	11
I do not believe	18	through trial and error	11	**would be able to**	22	**do not agree with**	11
if you do not	18	to keep up with	11	than to be the	21	I had to go	11
they are able to	18	**want to be a**	11	there are two reasons	21	in conclusion I think	11
you are able to	18	a wide variety of	10	first of all there	20	**in my opinion the**	11
a great way to	17	better than the other	10	thing I want to	20	as a matter of fact	11
a lot of the	17	from a young age	10	**one of the biggest**	19	is much better than	11
at my high school	17	go hand in hand	10	**the end of the**	19	is not good for	11
be one of the	17	I wish I could	10	**is a lot of**	19	it is obvious that	11
in the real world	17	I would love to	10	because of these reasons	18	the one of the	11
there are some things	17	**in my high school**	10	I am sure that	18	**this is not the**	11

APPENDIX A

Continued (Shin, 2018)

Native corpus (146 types, 2783 tokens)				Learner corpus (156 types, 3434 tokens)			
for example if you	16	in order to be	10	it can be a	18	to solve the problem	11
have the ability to	16	in order to get	10	**it is important to**	18	what they want to	11
there are a lot of	16	it comes down to	10	many people think that	18	**I believe that it**	11
it is easier to	16	may be able to	10	**one of the best**	18	**all over the world**	10
it would be to	16	people in the world	10	some people say that	18	**due to the fact**	10
one of the best	16	**I believe that it**	10	students who want to	18	he or she could	10
there are many different	16	that you have to	10	is not easy to	18	in the case of	10
there are plenty of	16	to do the same	10	**do not want to**	17	it is good to	10
a large amount of	15	when it came to	10	**for the first time**	17	it is often said	10
do not want to	15	you have to be	10	in my case I	17	it would be a	10
if I had to	15	studies have shown that	10	**in the middle of**	17	**not be able to**	10
in many different ways	15	**it is true that**	10	**it would be the**	16	so I agree with	10
in my opinion the	15			than to be a	16	**the best way to**	10
in the United States	15			the environment of my	16	the person who is	10
the world around us	15			there are several reasons	16	the reason is that	10
the world we live in	15			there are some reasons	16	**there are many different**	10
there is so much	15			therefore I want to	16	**there are many ways**	10
we are able to	15			is very important for	16	there are much more	10
as well as the	14			first reason is that	16	**there are some things**	10
easier for me to	14			for this reason I	15	who are good at	10
have a lot to	14			the problem is that	15	**I believe that the**	10
I feel as though	14			to take care of	15	**in my opinion I**	10

Applicable levels: College

Yu Kyoung Shin
Assistant Professor
Department of Global Studies
Hallym University
1 Hallymdaehak-gil, Chuncheon,
Gangwon-do 24252, Korea.
Email: yshin@hallym.ac.kr

Received on June 1, 2018
Reviewed on July 17, 2018
Revised version received on August 10, 2018