THE JOURNAL OF ASIA TEFL
Vol. 16, No. 2, Summer 2019, 516-530
http://dx.doi.org/10.18823/asiatefl.2019.16.2.5.516

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Syntactic Complexity of Recurrent Multiword Sequences in the Writings of Published Authors and L1 and L2 English Apprentice Writers

Yu Kyoung Shin
Hallym University, Korea

Huiseong Choi
Hallym University, Korea

Donghwan Kim
Hallym University, Korea

Seo-jeong Ko
Hallym University, Korea

Hyemin Yoo
Hallym University, Korea

Hyungjoon Yoo
Hallym University, Korea

Junsu Yoon
Hallym University, Korea

Isaiah WonHo Yoo
Sogang University, Korea

While previous studies have documented the internal structures of English lexical bundles (LBs) in a range of academic genres, how such fragmentary phrases are used in context has been given little attention. The current study addresses this gap by investigating to what extent three different writer groups, i.e., L1-English freshman students, L1-Korean freshman students, and published research article authors, integrate LBs into their English academic writing. The study first categorizes the LBs by their three main structures: VP-, NP-, and PP-based. It then focuses on the syntactic roles of the bundles in context, and plots these roles onto the developmental stages of grammatical complexity proposed by Biber, Gray, and Poonpon (2011). The results show a progressive sequence of native and nonnative apprentice writers' use of LBs toward being proficient academic writers, while also demonstrating that some features are unique to each group. The study extends the structural analysis of LBs in the literature by providing new information on the grammatical status of LBs in context and in relation to their uses by different groups of writers.

Keywords: syntactic complexity, syntactic roles, lexical bundles, research articles, argumentative essays, novice academic writers

Introduction

Formulaic language has attracted a great deal of attention from researchers in the last five decades. Pawley and Syder (1983) argued that formulaic language is an essential element of nativelike linguistic knowledge, which includes the ability to use routine sequences appropriately in context, and that the use of such sequences affects how the English of non-native speakers is perceived. Sinclair (1991) posited that texts are largely composed of "the occurrence of common words in common patterns, or in slight variants of those common patterns" (p. 108), challenging the conventional concept that it is grammar that determines lexical choice. Later work further suggested that formulaic language contributes to communicative efficiency and fluency in language production (e.g., Ellis, Simpson-Vlach, & Maynard, 2008; Hyland, 2012; Schmitt, 2004), because a whole string of words may be processed and used as a holistic unit rather than as discrete words (e.g., Nekrasova, 2009; Tremblay, Derwing, Libben, & Westbury, 2011; Wray, 2008).

One special type of formulaic language is lexical bundles (LBs), which are groups of three or more words that frequently recur in a genre (Biber, Johansson, Leech, Conrad, & Finegan, 1999). Examples of LBs in academic writing are expressions such as *in addition to the, is one of the*, and *to the development of.* Many researchers have utilized lexical bundles as a means of comparing first language (L1) and second language (L2) academic writing (e.g., Ädel & Erman, 2012; Bychkovska & Lee, 2017; Chen & Baker, 2010; Hyland, 2008a, 2008b; Kwon & Lee, 2014; Paquot, 2017; Ping, 2009; Qin, 2014; Salazar, 2014; Shin, 2019; Wei & Lei, 2011; Wood & Appel, 2014). These accumulated studies provide useful insights into the types and the internal structures of LBs used by L1 and L2 academic writers in a range of academic genres. However, how such fragmentary phrases are used in context has been given little attention, although lexical bundles are rarely complete grammatical structures in themselves, instead forming parts of larger structures. Shin (2018) addresses this issue by investigating LBs in terms of their syntactic roles such as adverbials, complements, and noun modifiers and also looks at the structural environments of LBs by examining the frequent co-structures of the bundles. Her study suggests that a close scrutiny of the syntactic roles of bundles, including the different roles that the same LB structure can play in sentences, can provide a more detailed picture of how different groups of writers use LBs in a given genre.

The current study focuses on the syntactic roles of lexical bundles used by three different writer groups (i.e., published authors, native English-speaking undergraduate students, and nonnative English-speaking undergraduate students) in academic prose. It then presents an attempt to map the use of bundles onto the developmental stages of syntactic complexity proposed by Biber, Gray, and Poonpon (2011). The results of the current study suggest a progressive sequence in which native and nonnative apprentice writers move toward becoming proficient academic writers, indicating that formulaic language could serve as a useful tool to measure the syntactic development of academic prose.

Literature Review

Formulaic Language

Formulaic language has long been a topic of research in applied linguistics (e.g., Allerton, 1984; Firth, 1957; Jespersen, 1924; Palmer, 1933). Firth, for example, pointed out that the meanings of words depend on the sequences in which they appear: "You shall know a word by the company it keeps" (1957, p. 11). While early work on formulaic language relied on the intuition of individual researchers to identify recurrent multiword sequences, advances in computer technology have made it possible to identify the sequences using more empirical methods, demonstrating the validity of these earlier proposals. In 1991, John Sinclair proposed two seminal concepts: the idiom principle, i.e. words do not stand in isolation but occur with each other to produce meaning, and the open-choice principle, i.e. words are selected to fill certain slots in a stock of prefabricated expressions. These concepts, along with the possibilities offered by computers, led to new developments in research on multiword sequences. For instance, concordances or n-gram identifiers enable researchers to search for multiword sequences. The accessibility of such technology has prompted extensive

corpus-based research on formulaic language (e.g., Ädel & Erman, 2012; Granger & Paquot, 2012; Kashiha, & Heng, 2013; Maswana, Kanamaru, & Tajino, 2013; Pérez-Llantada, 2014; Salazar, 2014; Shin, 2019; Shin, Cortes, & Yoo, 2018; Shin & Kim, 2017; Wray, 2002; Zipagan & Lee, 2018).

Individual researchers have developed their own methods for identifying multiword sequences, and these have varied depending on how the sequences are operationalized. For example, fixedness, idiomaticity, length, syntactic completeness, semantics, and frequency of sequences have all been used as criteria (Conrad & Biber, 2005). For this reason, although they can all be considered under the umbrella of phraseological units and formulaic language (e.g., Wray, 2002), many terms have been used to label different types of sequences, including "lexical phrase" (Nattinger & DeCarrico, 1992), "formulas" (Simpson-Vlach & Ellis, 2010), "clusters" (Scott, 1996), "n-grams" (Milton & Freeman, 1996), and "lexical bundles" (Biber et al., 1999), among others.

A widely used construct in research on such frequent word combinations is that of lexical bundles. The identification of lexical bundles is based solely on their frequency in a corpus, and in most cases, they are structurally incomplete, but closely related to certain types of structures. Several researchers (Biber et al., 1999; Salazar, 2014; Shin, 2019) have investigated the internal structures of lexical bundles, which vary according to genre. Biber et al. (1999), for example, grouped bundles into three main grammatical types based on their occurrence in the Longman Corpus of Written and Spoken English. They found that conversation consists of more clausal bundles such as verb phrase fragments (e.g., *I mean you know*) and dependent clause fragments (e.g., *that there is a*) while academic prose uses more phrasal bundles, including noun phrases (e.g., *one of the things*) or prepositional phrases (e.g., *at the same time*).

Syntactic Complexity in Relation to Lexical Bundles

A variety of methods have been used to measure syntactic complexity. For one, corpus-based research has documented structures that are typical of academic writing by proficient native writers (e.g., Biber et al., 1999; Biber & Gray, 2010; Biber et al., 2011; Gray, 2015). Biber and his colleagues have demonstrated that academic writing is characterized by structural compression, characterized by nominalizations and phrasal modifiers, while conversation is typically less compressed and more elaborated, characterized by the frequent use of subordinate clauses (Biber & Gray, 2010). These findings run counter to the traditional account that academic writing is grammatically complex with elaborated structures (e.g., Beers & Nagy, 2009, 2011; Hughes, 2005; Wolfe-Quintero, Inagaki, & Kim,1988). On the contrary, simple clause structures with complex phrases are very common in academic writing, e.g. "The distinctive effect [of the size [of the Asian population] [on income inequality] certainly *deserves* future research" (with head nouns underlined, verbs in *italics*, and post-modifying phrases in [brackets]) (Gray, 2015, p. 50).

With an increasing recognition that syntactic complexity is an important index of development in academic writing, a large number of L2 research studies have centered on the syntactic complexity of learners' academic writing (e.g., Lu, 2011; Lu & Ai, 2015; Mazgutova & Kormos, 2015; Norris & Ortega, 2009; Taguchi, Crawford, & Wetzel, 2013; Vyatkina, 2013; Weigle & Friginal, 2015), while very few studies, mostly targeting young learners, have investigated the syntactic complexity of L1 writings (e.g., Biber et al., 2011; Ravid & Berman, 2010; Staples, Egbert, Biber, & Gray, 2016). Taguchi et al. (2013), for instance, analyzed the syntactic complexity of low- and high-rated L2 argument essays at the clause and phrase levels. The two groups of essays used clausal features similarly but used phrasal features in distinctive ways, with high-rated essays more frequently employing noun phrase modification such as attributive adjectives and post-modifying phrases.

Biber et al. (2011) went further, proposing L2 developmental stages based on complexity features. They analyzed 28 grammatical features, found in published research articles and face-to-face conversation produced by native English speakers. Based on their observations of native speakers, they suggested a progressive sequence according to the features that L2 learners are expected to produce in their writings. Biber et al. proposed that L2 writers will first use the features typical of native-speaker conversation, followed by those common in research articles.

To date, no research has investigated the developmental stages of syntactic complexity in relation to

lexical bundles, although many researchers have documented the internal structures of lexical bundles. Because bundles are identified on the basis of frequency, they are mostly fragmented structures, embedded in larger structures (Shin, 2018). It should be possible, therefore, to observe how different populations embed bundles in context, and hence to plot their usage onto Biber et al.'s (2011) developmental stages. For example, each occurrence of the lexical bundle *there are lots of* found in both the native and nonnative corpora (see Section 3.1 for details of the corpus data) can first be structurally categorized depending on whether it occurs in a main verb phrase or in a dependent clause. Those in the dependent clause category can then be subcategorized by the various roles, which are associated with different developmental stages; for example, a FİNİTE COMPLEMENT CLAUSE CONTROLLED BY A VERB (e.g., *I believe [there are lots of things young people can teach older people]*), a FİNİTE ADVERBİAL CLAUSE (e.g., *In today's society, it's more important than anything to be open-minded and understanding [because there are lots of things being introduced daily]*), and a COMPLEMENT CLAUSE CONTROLLED BY A NOUN (e.g., *The current soccer field is just a thin layer of sand [where there are lots of rocks and other potential dangerous substances]*).1 The bundles and their syntactic roles used by L1 and L2 writers can then be discussed in terms of how they rank in the hypothesized developmental stages for complexity features.

Previous studies on syntactic complexity in academic writing, such as Taguchi et al.'s (2013), have shown that proficient writers use more phrasal complexity than do less proficient writers. They do not, however, provide sufficiently detailed information on the construction of phrasal/clausal structures in context, i.e. to what extent different writer groups integrate the structures into sentences. This is a crucial point because the main structural types of bundles identified in the previous studies (i.e., phrase and clause) can behave differently in a sentence depending on their syntactic roles, as the example of *there are lots of* demonstrates. LBs appear to have the potential to serve as a tool to compare the syntactic roles filled by lexical bundles in a target genre, as used by each population group. The following three research questions guided the current study:

1. What are the frequent four-word LBs used by native and nonnative freshman students and published authors? Do the LBs used by the three groups differ in terms of structural types?
2. To what extent does each group use the shared bundles (those found in all three corpora) in context?
3. To what extent do the syntactic roles of LBs used by each group conform to the stages of development for syntactic complexity features proposed by Biber et al. (2011)?

Method

Corpora

The learner corpus (LC, hereafter) in this study comprises argumentative essays from L1 Korean freshman students at a university in Korea. The data were collected from 2009 to 2012. A total of 4,214 students wrote essays as the placement test for freshman English classes, with a total of 1,018,524 words, as shown in Table 1. The students were asked to write an essay on one of eight writing topics in 50 minutes, one of which was "Do you agree or disagree with the following statement: People always learn from their mistakes?" Spelling errors were corrected, but errors such as *importantting* and *teached*, which involve incorrect grammatical information, were not corrected.

TABLE 1
Description of the Three Corpora

Corpora	Number of writings	Mean length of writings	Total corpus size
Learner Corpus (LC)	4,214	242 words	1,018,524 words
Native Corpus (NC)	1,479	345 words	509,516 words
Research article Corpus (RC)	148	6,871 words	1,016,882 words

[1] In each example, the whole phrase that fills the syntactic role under consideration is enclosed in square brackets.

The native corpus (NC, hereafter) consists of argumentative essays from L1 English freshman students at a university in the United States. The students were asked to write argumentative essays as a diagnostic test at the beginning of freshman composition courses in 2017 and 2018. The writing prompts and time constraints were identical to those for the essays in the learner corpus. The native corpus includes 1,414 essays, with a total of 509,516 words (see Table 1).

Lastly, the research article corpus (RC, hereafter) was used as the expert native corpus for this study; it consists of academic research articles published in the field of applied linguistics in international journals, amounting to about one million words (see Table 1). Note that the research articles are significantly longer on average (6,871 words) than the native and nonnative student writings; the total number of texts in each corpus differs accordingly.

For the LC and the RC, which contain approximately one million words each, raw frequencies were used without converting them to a normalized rate; the frequencies in the NC, which has about half a million words, were doubled to match the size of the other two corpora.

Lexical Bundles in Native and Nonnative Corpora

Following Hyland (2008a), the frequency threshold for a four-word bundle was set at 20 times in a million words, and the range threshold at a minimum of ten texts. *AntConc* (Anthony, 2014) was used to retrieve all the relevant data. In addition, the bundles identified in the LC and NC corpora were examined to exclude all "prompt" bundles, following Staples et al. (2013). If a bundle in a text was identical to a bundle used in the prompt for the writing task to which the text responded, it was removed from the data so that only non-prompt bundles were used for this study.

The identified bundles were then categorized in terms of their internal structures, following Biber, Conrad, and Cortes (2004), to examine how the usage of the bundles is different and/or similar in the three groups of writers. Three broad structural categories were employed: verb phrase (VP)/dependent clause fragments, noun phrase (NP), and prepositional phrase (PP) fragments. Verb phrase fragments refer to main verb fragments, for example, *That's one of the*. Dependent clause fragments include combinations with a dependent clause component with simple verb phrase fragments (*I want you to*) or dependent clause fragments starting with complementizers or subordinators such as *to be able to* and *what I want to*. Noun phrase fragments comprise those including NPs with *of*-phrase fragments (e.g., *the end of the*) and post-modifier fragments (*those of you who*), while prepositional phrase fragments consist of a preposition with an NP (*in the context of*).

Syntactic Roles and Hypothesized Developmental Stages for Complexity Features

Once all of the LBs identified in each corpus were thus categorized, the structural uses of the bundles were compared in order to examine how each group uses the shared bundles (those found in the three corpora) according to their syntactic role in the phrase or clauses that contain them.

The study then employed Biber et al.'s (2011) developmental stages to analyze the three groups of writers' use of *all* LBs in terms of syntactic roles. All the syntactic roles of the structures in which each LB was embedded were manually counted. This process required a close examination of each bundle in context, as LBs are generally fragmented phrases or clauses embedded in another structure. For instance, the lexical bundle *to go to the*, found in the learner corpus, was categorized according to syntactic roles, which correspond to different stages of complexity: a NON-FİNİTE COMPLEMENT CLAUSE (CC) CONTROLLED BY A COMMON VERB (e.g., *want*) – Stage 2 (e.g., *want to go to the*), a NON-FİNİTE CC CONTROLLED BY A WİDER SET OF VERBS – Stage 3 (e.g., *decided to go to the*), and a NON-FİNİTE CC CONTROLLED BY AN ADJECTİVE– Stage 4 (e.g., *dangerous to go to the*). At times, this bundle was used as an ADVERBİAL as in *Because many people take the big exam many times to go to the best college*; however, this type of structure is absent in Biber et al.'s developmental stages, and thus had to be excluded from the analysis.

Results

Lexical Bundles Identified in the Three Corpora

This section addresses the first research question regarding the LBs identified in the three corpora. The study initially found a total of over 300 four-word bundles in both native and nonnative student corpora (LC and NC), but it became apparent that many of the bundles (and even longer ones) came directly from the writing prompts. After such prompt bundles were excluded, 98 non-prompt bundles in the LC and 81 in the NC remained for the study, showing that the apprentice writers, regardless of their first language, draw on a large set of expressions.

As in Table 2, the nonnative student writer groups have approximately three times as many different types of lexical bundles as the research article authors. This pattern contradicts the findings from previous data-driven studies which showed that expert writers use a wider range of bundles than do novice academic writers and/or English learners (e.g., Ädel & Erman's 2012 report that published authors used 130 LBs and English learners 60).

TABLE 2
Types and Overall Frequency of Lexical Bundles Across Three Corpora

	LC	NC	RA
Type	98	81	34
Tokens	7015	6898	2168

Table 3 shows the distribution of structural types in the three corpora. As seen in the table, analogous patterns were observed in the two student writer groups (LC and NC), with almost the same proportion of the three internal structural types. In both corpora, VP-based bundles make up the greater part (about 65%) of all bundles; of the phrasal bundles, NP- and PP-based bundles occur at similar rates. The research article corpus (RC), on the other hand, contains a high proportion of phrasal bundles (about 80%), with considerably fewer VP-based bundles. The three corpora contain several bundles that do not fit the three types of structures: *as well as the* and *significant at the level* in the RA; some bundles involving the comparative (e.g., *more important than the*) in the NC; and colloquial expressions (*thank you for reading*) in the LC.

TABLE 3
Distribution of Main Structural Categories

Structural categories	Types		
	LC	NC	RC
NP-based – Phrasal	17.3% (17)	16% (13)	35.3% (12)
PP-based – Phrasal	16.3% (16)	14.8% (12)	44.1% (15)
VP-based – Clausal	64.3% (63)	66.7% (54)	14.7% (5)
Other	2% (2)	2.5% (2)	5.9% (2)
Total	100% (98)	100% (81)	100% (34)

Overall, the three corpora show divergent usage patterns. One point of difference is that the student writers (both natives and nonnatives) use clausal bundles much more frequently than do the expert writers. Biber and Gray (2010) found that subordinated clauses are commonly used in conversation, contributing to elaborated expressions, while academic writing is structurally compressed with the use of phrasal modifiers embedded in NPs. It is notable that the novice academic writers are less likely to incorporate LB types specific to academic prose, but instead heavily rely on types that are typically used in face-to-face conversation, such as *I would like to*, *there are lots of*, and *when it comes to*. The following examples demonstrate uses of *I would like to* found in the LC (1) and the NC (2), in response to the same writing prompt.

(1) Doing this would help the administration in managing the number of immigrants moving into the town. In relation to the other points, **I would like to** fully computerize the administration of the town so as to make administration streamlined and efficient. (LC)

(2) I admit asking my city to single handedly overcome racism is a stretch, but it is an important issue that **I would like to** see my city and my country make strides toward. (NC)

Syntactic Roles of Lexical Bundles

This section discusses the second research question, regarding how native and nonnative undergraduate students and expert writers use the same bundles (those shared by the three groups) in terms of syntactic roles.

TABLE 4
Syntactic Roles of Shared Bundles in the Three Corpora, with Numbers of Tokens and Percentages

LBs	Syntactic Roles	LC	NC	RC
on the other hand	adverbial	175 (100%)	240 (100%)	209 (100%)
in the case of	adverbial	13 (100%)	14 (100%)	131 (100%)
as well as the	adverbial	5 (100%)	30 (100%)	100 (100%)
in terms of	adverbial	2 (100%)	64 (100%)	90 (100%)
the results of the	subject	-	-	39 (45.3%)
	subject predicative	-	-	1 (1.2%)
	object	1 (100%)	-	32 (37.2%)
	complement in PP (adverbial)	-	-	5 (5.8%)
	complement in PP (post nominal modifier)	-	-	9 (10.5%)
the fact that the	subject	2 (22.2%)	2 (14.3%)	22 (29.7%)
	subject predicative	-	-	2 (2.7%)
	object	6 (66.7%)	4 (28.6%)	14 (18.9%)
	complement in PP (adverbial)	1 (11.1%)	8 (57.1%)	32 (43.2%)
	complement in PP (post nominal modifier)	-	-	4 (5.4%)
the rest of the	subject	1 (33.3%)	6 (13.6%)	23 (31.9%)
	subject predicative	1 (33.3%)	-	1 (1.4%)
	object	-	18 (40.9%)	5 (6.9%)
	complement in PP (adverbial)	1 (33.3%)	-	37 (51.4%)
	complement in PP (post nominal modifier)	-	2 (4.5%)	4 (5.6%)
	adverbial	-	18 (40.9%)	1 (1.4%)
	appositive	-	-	1 (1.4%)
on the basis of	adverbial	1 (100%)	-	69 (100%)
	complement in PP (adverbial)	-	4 (100%)	-
at the same time	adverbial	61 (100%)	54 (100%)	64 (100%)
as a result of	adverbial	9 (100%)	16 (100%)	63 (100%)
the size of the	subject	3 (50%)	2 (100%)	11 (18.3%)
	subject predicative	1 (16.7%)	-	2 (3.3%)
	object	1 (16.7%)	-	11 (18.3%)
	complement in PP (adverbial)	1 (16.7%)	-	29 (48.3%)
	complement in PP (post nominal modifier)	-	-	5 (8.3%)
	appositive	-	-	2 (3.3%)
at the end of	adverbial	13 (100%)	64 (100%)	42 (72.4%)
the value of the	post nominal modifier	-	-	16 (27.6%)
	subject	2 (40%)	4 (50%)	19 (33.9%)
	subject predicative	1 (20%)	-	3 (5.4%)
	object	2 (40%)	2 (25%)	16 (28.6%)

that there is a	complement in PP (adverbial)	-	-	2 (8.3%)
	complement in PP (post nominal modifier)	43(78.2%)	-	6 (25%)
	CC controlled by Verbs	-	7 (100%)	12 (50%)
	CC controlled by common Verbs	12(21.8%)	-	2 (8.3%)
	finite adverbial clause	-	-	2 (8.3%)
at the time of	adverbial	46 (85.2%)	1 (100%)	-
	post nominal modifier	7 (13%)	-	-
	appositive	1 (1.9%)	-	-
the end of the	subject	2 (9.5%)	-	-
	subject predicative	2 (9.5%)	2 (2.4%)	-
	object	17 (81%)	2 (2.4%)	1 (2.0%)
	complement in PP (adverbial)	-	76 (92.7%)	41 (80.4%)
the difference between the	complement in PP (post nominal modifier)	-	2 (2.4%)	9 (17.6%)
	subject	2 (66.6%)	4 (50%)	17 (36.2%)
	subject predicative	1 (33.3%)	-	4 (8.5%)
	object	-	2 (25%)	5 (10.6%)
	complement in PP (adverbial)	-	2 (25%)	15 (31.9%)
	complement in PP (post nominal modifier)			2 (4.3%)
to the extent that	appositive	-	-	4 (8.5%)
	adverbial	1 (100%)	8 (100%)	45 (100%)
in the form of	adverbial	-	6 (75%)	22 (52.4%)
	post nominal modifier	1 (100%)	2 (25%)	20 (47.6%)
to the fact that	adverbial	6 (100%)	70 (100%)	42 (100%)
it is important to	main verb fragment	33 (78.6%)	28 (70%)	38 (95%)
	CC controlled by verbs	-	8 (20%)	-
	finite complement clauses	6 (14.3%)	-	1 (2.5%)
	CC controlled by nouns	-	2 (5%)	1 (2.5%)
	Finite adverbial clauses	3 (7.1%)	2 (5%)	-
Total		385 (100%)	794 (100%)	1507 (100%)

As can be seen in Table 4, all the shared bundles except for one (*it is important to*) consist of noun phrases or prepositional phrases, which together make up 95.2% of the shared bundles. The total number of bundles shared by all three groups is rather small, at only 21. This is partly because the clausal bundles used by the undergraduate writers (native and nonnative alike) include colloquial and idiomatic expressions, which is not the case for the published authors. In addition, whereas the two groups of student writers shared many phrasal bundles, many of them are more typical of spoken genres than of academic prose, such as *for a long time*, *for the first time*, and *with that being said*. However, the 21 bundles shared by all three groups are typical of formulaic sequences, in line with previous findings in the literature (e.g., Biber et al., 1999, 2004; Chen & Baker, 2010; Salazar, 2014; Pan, Reppen, & Biber, 2016).

The syntactic roles of the NP-based bundles are presented in Figure 1. One finding that stands out is that the usage of the English learners (i.e., the nonnative English-speaking student writers) exhibits patterns that diverge from those shown by the other two groups (native student writers and published authors).

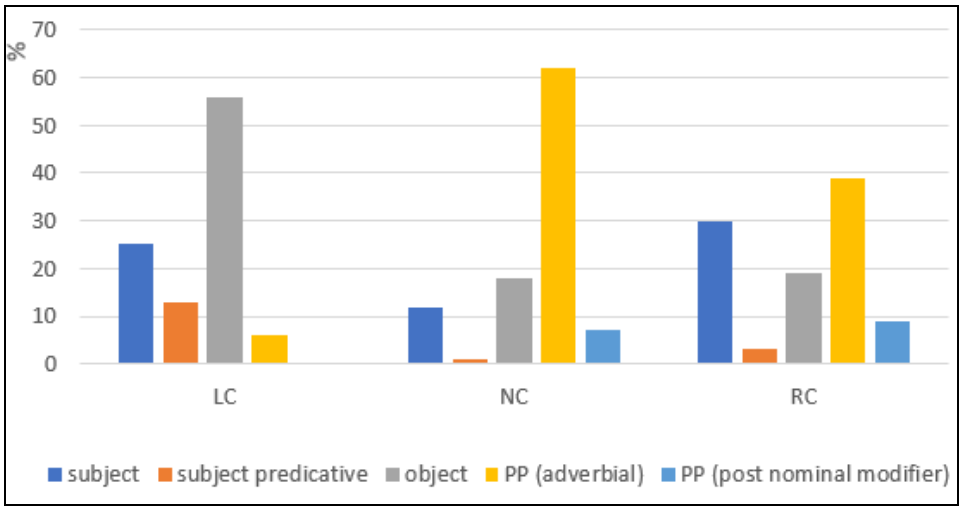


Figure 1. Distribution of syntactic roles of NP bundles in the three corpora.

As shown in Figure 1, the English learners mostly used NP bundles as subjects or objects, which together account for about 94%: objects (56%), subjects (25%), and subject predicatives (12.5%). This distribution exhibits greater differences in how NP bundles are used, compared to the other two groups. In both the NC and RC, NP bundles appear most frequently as complements of PPs to function as adverbials. The native and nonnative student writer groups, in particular, displayed extreme differences. The native student writers overused this role, at about 60%, whereas the learners underused it, at only 6%. In addition, the native students and published authors, but not the English learners, sometimes embedded NP bundles in PPs functioning as postmodifiers, albeit infrequently (NC: 7%, RC: 9%).

The learners were also found to frequently use NP bundles in the role of objects (56%), while the other two groups did so only infrequently, at very similar rates (NC: 18%, RC: 19%). Moreover, the learners were likely to use phrasal bundles in subject predicatives (13%), placing them after *be*-verbs, more often than did the published authors (3%) and the native students (1%).

The following examples illustrate the different syntactic roles of the same LB in context. One NP bundle, *the size of the*, plays the role of subject, subject predicative, and object in (3), (4), and (5), respectively. Examples of a complement in a PP functioning as an adverbial and post nominal modifier appear in (6) and (7). All the examples were found in the LC, except for (7), which is from RC.

(3) And, **the size of the** letters are really small so it takes pretty long time to read all of the articles. (LC)
(4) However, on the other hand, the problem is **the size of the** school. (LC)
(5) If I could make one important change in a school which I attended, I want to **change the size of the** school. (LC)
(6) If someone is coming to Seoul for the first time, they might be surprised at **the size of the** city. (LC)
(7) Whilst there is perhaps some evidence of a mild reduction in the range of mean absolute returns on the fourth and fifth days of the 5-day correlogram, there is little evidence of any decay in the size of autocorrelations over the week, or of any substantial increase in **the size of the** daily autocorrelations at the fifth or weekly frequency, as has been reported for other markets. (RC)

Regarding the syntactic roles of PP bundles, one small but noticeable difference across the groups is the use of such bundles as post nominal modifiers by both the native students and the expert writers (NC: 1.8%, RC: 6%), but not by the English learners. The following examples of *at the end of* illustrate the roles of PPs: as a post nominal modifier following a nominal phrase (i.e., *Conclusions* in the RC and *the light* in the NC) in (8) and (9) and as an adverbial in the LC in (10).

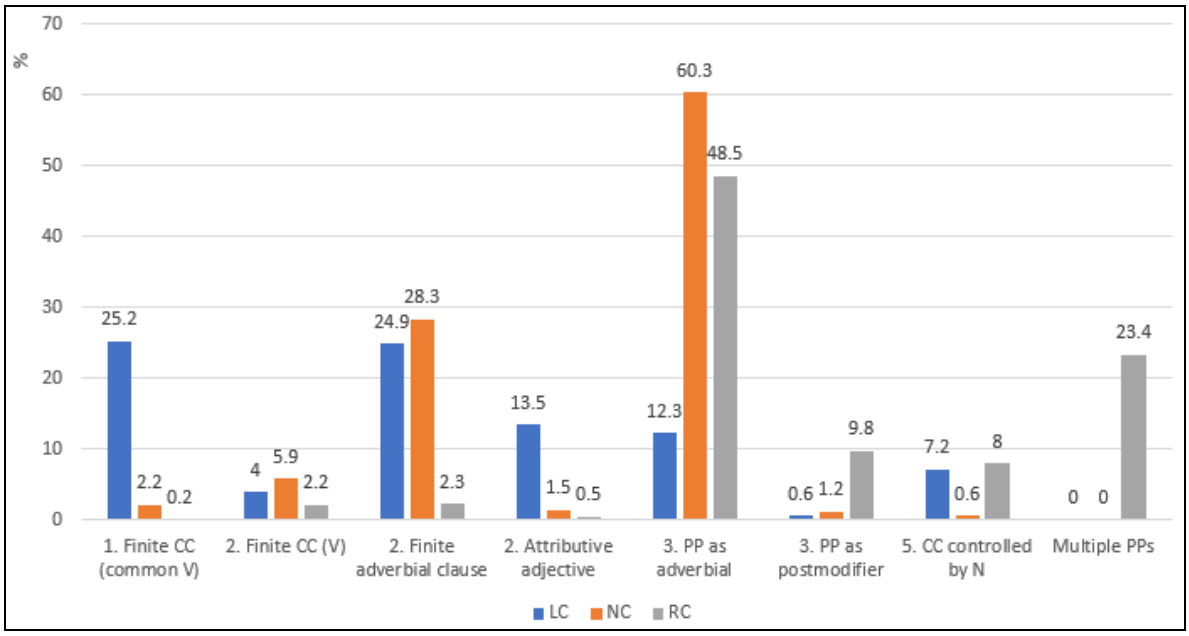
(8) The two other theses were unusual in that one did not have a chapter which played a concluding role, and the other had Conclusions **at the end of** each Part but no overall conclusion to the thesis as a whole. (RC)
(9) The change that I made at Lovejoy High School is to simply see the light **at the end of** the tunnel. (NC)
(10) There are signs that tell what's **at the end of** the road, but too far apart, making it hard to check each one. (LC)

Almost all the previous studies on LBs have focused on the structural type of LBs *per se* (i.e., NP, PP, and VP structures). These studies have consistently found that native and/or expert writers favor phrasal bundles, while nonnative and/or apprentice writers often use clausal bundles. As noted above, however, bundles are fragmentary structures solely based on frequency, and thus mainly occur across structures (Shin, 2018). In other words, depending on the larger structures in which LBs are embedded (i.e., syntactic roles), the bundles serve different grammatical functions in context.

Recall that the native and nonnative student writers produced bundles in strikingly similar structures (see Table 3). The current study has extended the analysis of bundles' use in relation to grammatical structure by showing how differently bundles are constructed in context. The findings of this section, by examining the shared bundles, revealed distinctive features specific to English learners (e.g., their preference for bundles in the role of object, but rarely embedded in PPs).

Hypothesized Developmental Stages for Complexity Features

For the last question, the syntactic roles of *all* the bundles (not limited to the shared bundles) used by each group were plotted onto Biber et al.'s (2011) developmental stages. As shown in Figure 2, this study's three corpora contained most of the tokens of LBs in eight structures at four stages, out of Biber et al.'s 28 structures at five stages, and thus the percentages for each group do not necessarily add up to 100%.



Notes. The number preceding each structure refers to the stage; CC = complement clause.

Figure 2. Lexical bundles and developmental stages of complexity features.

Overall, the use of LBs plotted onto the developmental stages displayed patterns specific to each group and/or to two groups and further showed where each of the three groups is located in the developmental sequence. First, one fourth of the LBs used by the learners were embedded in a structure ranked at Stage 1: a

FİNİTE COMPLEMENT CLAUSE CONTROLLED BY AN EXTREMELY COMMON VERB (e.g., *think*, *know*, *say*). In this structure, the learners overwhelmingly used the verb *think*, suggesting that this is a feature unique to English learners. Both the native students and the expert writers rarely incorporated LBs into this type of structure.

Another notable difference is in their production of LBs in a structure ranked at Stage 2: a FİNİTE ADVERBİAL CLAUSE. Both native and nonnative student groups very often used bundles in this structure (about 25% each), whereas the expert writers did so infrequently (2.2%). Another Stage 2 structure, an ATTRİBUTİVE ADJECTİVE, was also much more common in the learner corpus (LC: 13.5%,). Many of the L2 uses of this structure involved bundles with a phrasal quantifier followed by a noun (e.g., *a lot of information*).

In the structures at Stage 3, however, the native students and expert writers used LBs more frequently than the learners. The biggest difference between these groups appeared in the use of the structure PP AS ADVERBİAL: About half of the bundles used by expert writers are embedded in this structure (48.5%), as are even more of those used by native students (60.3%), compared with less than a quarter of the LBs used by L2 learners (12.3%). At Stage 3, an OF-PHRASE USED AS AN POSTMODİFİER appeared much more frequently in the RC corpus (9.8%) than in the native and nonnative student corpora. In fact, most of the LBs used by the expert writers were in prepositional phrases, *of*-phrases in particular.

At the last stage, the learners and expert writers used LBs in a COMPLEMENT CLAUSE CONTROLLED BY A NOUN at similar rates (LC: 7.2%, RC: 8%), while the native students rarely did so. Biber et al. (2011) claimed that native speakers use this kind of clause much more frequently in academic writing than in conversation, and hypothesized that L2 learners are less likely to use it. The present study, however, found that the use of LBs in this structure was observed in the LC and RC, but was almost absent in the native student writing.

The structure ranked the highest in Stage 5 is EXTENSİVE PHRASAL EMBEDDİNG İN AN NP (multiple PPs as postmodifiers). While the expert writers used LBs in this structure frequently (23.4%), both the native and the nonnative student writers never did. The expert writers used bundles containing two prepositions generally consisting of a PP and the beginning of another PP (e.g., *in the case of*).

In sum, of particular interest is the finding that learners frequently use elaborated structures such as finite complement clauses controlled by verbs and finite adverbial clauses (together, 67.6%), which Biber et al. placed at Stages 1 and 2, while expert writers are most likely to use LBs in compressed structures typical of academic prose, including PPs as postmodifiers and multiple PPs (together, over 90%), which are ranked mostly at Stages 3 and 5. Given the consistent findings that native student writers have a head start over nonnative student writers (e.g., Ping, 2009; Salazar, 2014; Bychkovska & Lee, 2017), one might expect that the native writers would exhibit distinctive differences from the English learners. As discussed above, however, the results of this study suggest a continuum, on which the native students overlap with both the English learners (e.g., at stage 2 in terms of finite complement clause use) and the expert authors (e.g., at stage 3 in terms of the use of PPs as adverbials).

Interestingly enough, certain grammatical features (e.g., COMPLEMENT CLAUSE CONTROLLED BY A NOUN) that are common in academic prose, and are ranked highest in Biber et al.'s stages, are not manifested in the native student writing, but are in the learner writing. This finding may be partly attributable to the explicit formal instruction that the learners had previously received in EFL contexts, thereby making them more aware of specific structures and expressions typical of academic prose.

Overall, the developmental stages proposed by Biber et al. were found to be applicable to the use of LBs, demonstrating a progressive sequence of the structural patterns in which LBs are embedded when they are used by different groups of writers.

Conclusion

The current study investigated the relationship between LBs and syntactic complexity, revealing distinctive features specific to the use of LBs in academic written texts by different groups of writers. Because LBs are simply based on frequency, most of them are structurally incomplete. Consequently, even the same bundles can take multiple syntactic roles in context, which enables us to observe how the bundles

used by each group are related to syntactic complexity.

The first part of the study, following standard procedures in the internal structures of LBs, categorized the bundles' three structures (i.e., VP-, NP-, and PP-based). The findings demonstrated analogous patterns, with a very similar proportion of the three internal structural types between the English learners and the native students. Next, it focused on the syntactic roles of shared LBs in context, showing the divergent uses of the LBs, especially between the English learners and the other two groups (the native students and expert writers). The roles of all the bundles used by each group were then mapped onto the developmental stages proposed by Biber and his colleagues (2011). The findings showed that the novice writers (native and nonnative alike) were on a continuum with proficient writers. These results suggest that future research that delves more deeply into the association of syntactic complexity and formulaic sequences will provide a window onto writers' development of syntactic complexity.

The current findings, however, cannot be considered conclusive, given the different academic genres used in the corpora, i.e. argumentative essays in the LC and NC were compared to published research articles in the RC. Future analyses with parallel corpora controlled for specific academic genres as well as writing prompts will demonstrate whether the differences found in this study are actual differences in the uses of LBs unique to distinct populations of English writers. Furthermore, it would be interesting to examine LBs used by learners of different proficiencies in a future study in order to observe the relationships between LB usage and developmental sequences in more detail. One surprising finding of this study was that the learners and the native speakers used complement clauses controlled by nouns (Stage 5) at similar rates.

Given that it may be challenging for novice student writers and English learners to integrate multiword sequences into their writing, they would benefit from instruction on how LBs are integrated in context and on the structures that frequently co-occur with specific LBs. Instructors could utilize certain structure types that both native and nonnative writers strongly favor in the construction of LBs, linking them to linguistic complexity features appropriate to the target genre. In particular, the findings reported in this study demonstrated that learners frequently employ LBs in subjects and subject predicatives. Teachers should help learners to incorporate formulaic sequences in different text positions by modeling how the sequences can be combined with other types of structures or by providing essay samples with the same bundles taking different syntactic roles along with co-occurring structures.

Acknowledgement

This research was supported by Hallym University Research Fund, 2018 (HRF-201812-015).

The Authors

Yu Kyoung Shin (first author) is Assistant professor in the School of Global Studies at Hallym University. Her research interests include corpus analysis of academic written registers and disciplinary variation, particularly for applications to L2 language learning and teaching. Her work has recently been published in journals such as the *Journal of English for Academic Purposes*, the *Journal of Second Language Writing*, *Language Teaching Research*, *System*, and *Language and Intercultural Communication.*

School of Global Studies
Hallym University
Chuncheon, 24252, Korea
Tel: +82 332482486
Fax: +82 332482485
Email: yshin@hallym.ac.kr

Huiseong Choi, *Donghwan Kim*, *Seo-jeong Ko*, *Hyemin Yoo*, *Hyungjoon Yoo*, and *Junsu Yoon* (co-authors) are undergraduate students in the School of Global Studies at Hallym University.

School of Global Studies
Hallym University
Chuncheon 24252, Korea
Email: heesung4548@naver.com, dhk1611@naver.com, revibk_16@naver.com, yhm960729@naver.com, dbgudwbs31@gmail.com, jason930621@gmail.com

Isaiah WonHo Yoo (corresponding author) is Professor in the Department of English Literature & Linguistics at Sogang University. His primary research focuses on how corpus linguistics informs language pedagogy. His recent publications have appeared in *Applied Linguistics*, the *Journal of Second Language Writing*, *Language Acquisition*, and *Linguistic Inquiry*.

Department of English
Sogang University
Seoul, 04107, Korea
Tel: +82 27058340
Fax: +82 27150705
Email: iyoo@sogang.ac.kr

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.

Allerton, D. (1984). Three (or four) levels of word co-occurrence restriction. *Lingua, 63*, 17-40.

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

Beers, S., & Nagy, W. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing, 22*, 185-200.

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., 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.

Conrad, S., & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. In W. Teubert & M. Mahlberg (Eds.), *The corpus approach to lexicography. Thematic part of lexicographica. international Jahrbuch fuer Lexikographie, 20* (pp. 56-71). Berlin: De Gruyter.

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.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955. In J. Firth (Ed.), *Studies in linguistic analysis* (pp. 1-32). Oxford: Blackwell.

Granger, S., & Paquot, M. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32*, 130-149.

Gray, B. (2015). On the complexity of academic writing: Disciplinary variation and structural complexity. In V. Cortes & E. Csomay (Eds.), *Corpus-based research in Applied linguistics: Studies in honor of Doug Biber* (pp. 49-77). Amsterdam: John Benjamins.

Hughes, R. (2005). *English in speech and writing: Investigating language and literature*. New York, NY: Routledge.

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.

Jespersen, O. (1924). *The philosophy of grammar*. London: George Allen & Unwin.

Kashiha, H., & Heng, C. (2013). An exploration of lexical bundles in academic lectures: Examples from hard and soft sciences. *The Journal of Asia* TEFL, *10*(4), 133-161.

Kwon, Y., & Lee, E. (2014). Lexical bundles in the Korean EFL teacher talk corpus: A comparison between non-native and native English teachers. *The Journal of Asia* TEFL, *11*(3), 73-103.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36-62.

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing, 29*, 16-27.

Maswana, S., Kanamaru, T., & Tajino, A. (2013). Analyzing the journal corpus data on English expressions across disciplines. *The Journal of Asia* TEFL, *10*(4), 71-96.

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for academic purposes programme. *Journal of Second Language Writing, 29*, 3-15.

Milton, J., & Freeman, R. (1996). Lexical variation in the writing of Chinese learners of English. In C. E. Percy, Ch. F. Meyer, & I. Lancashire (Eds.), *Synchronic corpus linguistics*: *Papers from the sixteenth International Conference on English Language Research on Computerized Corpora* (pp. 121-131). Atlanta: Rodopi.

Nattinger, J., & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford: OUP.

Nekrasova. T. (2009). English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning, 59*(3), 647-686.

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*, 555-578.

Palmer, H. (1933). *Second interim report on English collocations*. Tokyo: Kaitakusha.

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.

Pawley, A., & Syder, H. (1983). Two puzzles for linguistic theory: Native-like selection and native-like fluency. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191-227). London: Longman.

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.

Ping, P. (2009). A study of the use of four-word lexical bundles in argumentative essays by Chinese English majors: A comparative study based on WECCL and LOCNESS. *CELEA Journal, 32*(3), 25-45.

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

Ravid, D., & Berman, R. (2010). Developing noun phrase complexity at school age: A text-embedded cross-linguistic analysis. *First Language, 30*(3), 3-26.

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

Scott, M. (1996). *Wordsmith Tools 4*. Oxford: Oxford University Press.

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

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list (AFL). *Applied Linguistics, 31*(4), 487-512. doi: 10.1093/applin/amp058

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press, Oxford, UK.

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

Shin, Y. (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, Y., Cortes, V., & 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.

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*, 214-225.

Staples, 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.

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly, 47*, 420-430.

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.

Vyatkina, N. (2013). Specific syntactic complexity: Developmental profiling or individuals based on an annotated learner corpus. *The Modern Language Journal, 97*, 11-30.

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

Weigle, S., & Friginal, E. (2015). Linguistic dimensions of impromptu test essays compared with successful student disciplinary writing: Effects of language background, topic, and L2 proficiency. *Journal of English for Academic Purposes, 18*, 25-39.

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.

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

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

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

Zipagan, M., & Lee, L. (2018). Korean English learners' use of lexical bundles in speaking. *The Journal of Asia TEFL, 15*(2), 276-291.