THE JOURNAL OF ASIA TEFL
Vol. 14, No. 4, Winter 2017, 703-716
http://dx.doi.org/10.18823/asiatefl.2017.14.4.8.703

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Functional Distribution of Lexical Bundle in Native and Non-Native Students' Argumentative Writing

Daehyeon Nam
Ulsan National Institute of Science and Technology, Korea

Although the idea of collocation has been the core of the Firthian linguistic approach, only recently did language researchers and practitioners pay close attention to recurrent multi-word combinations or lexical bundles. Unlike the studies of roles of collocation in a language, the research of lexical bundles has shown that they are crucial building blocks of discourse and register in academic disciplines. Given the fact that the role of lexical bundles in the writings of academic community has become a subject of ESL/EFL, the current study investigates the functional distribution of lexical bundles between the argumentative writing corpora contributed by Korean and American college students. The functional lexical bundle types and tokens are compared regarding the distribution and frequency. A series of statistical analyses demonstrated that, in the Korean students' argumentative writing, stance lexical bundle and discourse organizers are more frequently used. On the other hand, the American students' writing includes more referential expressions. These findings suggest that, when aiming for a proficient and advanced writing level, Korean college students need to be frequently exposed to lexical bundles and explicitly taught with the discourse functions in the academic writing genre. This paper also introduces and discusses pedagogical implications and future research suggestions.

Keywords: lexical bundle, functional distribution, corpus, argumentative writing

Introduction

Collocation, a linguistic tendency of lexical co-occurrence, has been a widely recognized language pattern among linguists and language educators (Firth, 1957; Sinclair, 1991). In language studies, therefore, the frequent occurrence of consecutive word units has been one of the important subjects in the fields (Biber, Johansson, Leach, Conrad, & Finegan, 1999; Chen & Baker, 2010; Cortes, 2004; Hyland, 2008a). From the empirical point of view on recurrent word combinations, the use of corpora and its methodological values have been acknowledged to capture and analyze language phenomena in given contexts or registers. One of the immediate reasons why linguists and language practitioners adopt the methodology is because it allows them to manage a huge amount of text data to investigate the collocations of two-word sequences and even longer word sequences: three-word, four-word, five-word, or six-word sequences. This type of word sequences is considered to be units of language and often called *lexical bundles* (Biber et al., 1999; Biber & Barbieri, 2007; Chen & Baker, 2010).

Different researchers have explored the lexical bundle using various terms. For example, cluster, a similar linguistic structure stemming from the concept of collocation, refers to a contiguous sequence of a certain number of words without considering its grammatical structure (Scott & Tribble, 2006). A number of studies have also focused on multi-word expressions labeled as *lexicalized sentence stem* (Pawley &

Syder, 1983), *lexical phrase* (Nattinger & DeCarrico, 1992), *formulaic sequence* (Wray, 2006), *phraseology* (Cowie, 1998), *chunks* (De Cock, 1998), or *n-grams* in computational linguistics and computer science (Manning & Schütze, 2001). Structurally, these terms ultimately refer to contiguous word sequences retrieved from corpora with specified frequency and distribution criteria. With all the different names and terms related to the word sequence, the lexical bundle is recognized as conventional grammar structures and discourse functions used and accepted by the speakers of a language within certain contexts (Chen & Baker, 2010). For example, Pawley and Syder (1983), one of the earliest explorations of the recurrent word combinations, recognize native-like expressions among the grammatically correct phrases of ordinary mature native speakers; these expressions are wholly or largely fixed linguistic units. Thus, for native speakers, lexical bundles seem to be ways to produce L1 and sound like native speakers whereas, for L2 learners, mastering the lexical bundles is one of the hurdles that they should overcome to sound like native speakers.

Biber, Conrad, and Cortes (2004) also discussed the lexical bundles being one of the basic linguistic units constructing discourse in writing. They argued that the lexical bundles differ from other similar ideas of word combinations, such as traditional formulaic expressions, because the lexical bundles are strictly defined by the frequency of the occurrence without arbitrary linguistic considerations. The investigation of the lexical bundle functions in L1 and L2 corpora is, therefore, expected to indicate how language learners construct their ideas in the discourse of writing compared to native speakers' construction.

Within corpus linguistics, the study of L2 English corpus is one of the fastest growing areas as English has become a lingua franca in many registers. For example, Granger and her colleagues built the International Corpus of Learner English (ICLE; Granger, Dagneaux, Meunier & Paquot, 2009) for researchers and educators to compare and contrast the writing patterns of L2 English learners with different L1 backgrounds. Given the recent linguistic interests in lexical bundles, the methodological development of corpus study, and the emerging role of L2 English in academic purposes, the exploration of L2 lexical bundles would contribute to explaining how linguistic units behave and are distributed according to certain L1 backgrounds.

Literature Review

While grammars are constructed on the basis of the *open choice* of lexical items, there is another important principle about word combination, so-called *idiom principle*: a large number of semi-prefabricated word combinations that are available within a certain register (Sinclair, 1991). As one of the key concepts of idiomaticity, the principle explains collocation as word sequences of two or more words within a short span of words in a text. The linguistic concept of collocation has empirically explained how a word in a register prefers certain word combinations. Although this insightful idea of collocation may not capture the differences in terms of whether certain word sequences are idiomatic or not (e.g., *hit the road* vs. *thank you very much*), the concept of lexical bundle, on the other hand, allows researchers to effectively examine the behavior of idiomatic and non-idiomatic recurrent multi-word sequences in a given register. In the exhaustive corpus-based grammar of spoken and written English, Biber et al. (1999), explained that lexical bundles are recurrent multi-word sequences based on a statistical tendency to appear together regardless of their idiomaticity. In this respect, the lexical bundle approach to spoken and written language provides valuable pieces of information to explain how a language is built in certain genres, disciplines, and language communities.

As computer technology advances, language researchers and educators are able to explore a large amount of linguistic text data in efficient ways. Biber et al. (1999) carried out a large-scale corpus study to find and redefine grammar in academic prose and conversation. A part of the study explored what types of lexical bundles exist and how the lexical bundles are distributed in different registers of English. Subsequent studies of the lexical bundle have focused on the usage patterns of lexical bundles across

registers, genres, and disciplines as well as the native/non-native language users. Of the studies, lexical bundle research in academic disciplines has been actively conducted especially in university registers (Biber, 2006a, 2006b; Biber et al., 2004), the general academic genre (Chen & Baker, 2001; Durrant, 2015; Pérez-Llantada, 2014), general academic writing among native and non-native scientific writers (Salazar, 2014), history and biology (Cortes, 2004), and telecommunications (Pan, Reppen, & Biber, 2016).

Biber et al. (1999) defined lexical bundles as "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (p. 990). In other words, lexical bundles are commonly occurring word combinations in natural language. The lexical bundles are then identified empirically as word combinations that recur most commonly in a register. To have word combinations qualified for lexical bundles, the combinations must frequently recur in a register. Theoretically, two-word, three-word, four-word, or longer lexical bundles are possible because of the frequency-based operational definition of the lexical bundle. However, five-word or longer bundles are generally less common. Therefore, most of the lexical bundle studies, including Biber et al. (1999), have focused on four-word lexical bundles (e.g., Chen & Baker, 2010; Cortes, 2004; Hyland, 2008a).

Given that one of the characteristics of lexical bundles is defined as their recurrence, it is crucial to set a criterion of frequent lexical bundles in a given text. In a corpus linguistic analysis, frequency data should be controlled cautiously to ensure a principled and governed statistical analysis. To the researchers' dismay, according to existing literature, there is no consensus among individual studies regarding the cut-off points of legitimate lexical bundle recurrence. In fact, the actual frequency of lexical bundle cut-off arbitrarily ranges from 20 times per million words (Cortes, 2004; Hyland, 2008a; Reppen, 2009; Wei & Lei, 2011) to 25 times per million words (Chen & Baker, 2010) or even 40 times per million words (Biber & Barbieri, 2007). For smaller corpora, such as spoken corpora, a raw cut-off frequency can be used from two to 10 times (De Cock, 1998). Another criterion for the lexical bundle selection should be the number of occurring lexical bundles in texts in a corpus. Studies have suggested that lexical bundles occur across three to five texts (Biber & Barbieri, 2007) or 10% of texts (Hyland, 2008a). The last lexical bundle selection issue should be the length of the lexical bundles, usually ranging from two- to six-word combinations. Of the combinations, four-word lexical bundles are considered to be the most researchable bundles because (1) four-word bundles encompass three-word bundles and the bundles occur more frequently than five-word bundles (Cortes, 2004; Hyland, 2008b) and, therefore, (2) the number of four-word bundles can be a manageable size for manual observation and categorization (Chen & Baker, 2010).

Although manageable sizes of lexical bundles are useful for research purposes, the lexical bundles also need to be customized for pedagogical purposes. Language educators and practitioners, such as Biber et al. (2004) and Hyland (2008a), have suggested functionally categorized lexical bundles to determine the lexical bundles more relevant for learners. Considering the significance and usefulness of lexical bundles for enhancing professional proficiency in academic writing, it is crucial to determine and categorize discourse functions of the lexical bundles to better fit classrooms and academic writing purposes (Römer, 2010).

The lexical bundles occur in different distributional patterns in terms of the nature of registers. Biber et al. (1999) reported interesting findings of the lexical bundles in academic prose and conversation registers. They found that 15% of the lexical bundle in the conversation register while 5% of the lexical bundles in the academic prose register can be considered complete grammatical units. In the conversation register, approximately 50% of the lexical bundles begin with a personal pronoun followed by a verb phrase (VP; e.g., *I don't know what*). On the other hand, 60% in the lexical bundles in academic prose are noun phrases (NPs) or prepositional phrases (e.g., *the nature of, as a result of*).

Academic writing has been one of the most researched topics in English education. Studies have revealed that language in uses is characterized by a recurrence of fixed and semi-fixed multi-word sequences or formulaic patterns in academic disciplines (e.g., Byrd & Coxhead, 2010; for native academic writing). Non-native writers' lexical bundle usages have also been studied (e.g., Hyland, 2008a,

2008b). Student writers struggle to make their texts fluent and assured to the readers in their disciplines. For example, Hyland (2008b) argued that the lexical bundles shape meanings and coherence in an academic text. For this reason, student writers are challenged to make their writing sound natural and to be accepted in their academic communities. After analyzing a corpus of professional research articles, doctoral dissertations, and master's theses of 3.5 million words in four disciplines the researcher concluded that each genre presents different and distinctive patterns of lexical bundles. The findings of these learner corpora suggest more realistic models of writing for students, improved pedagogy from language learning activities to curriculum, and the materials development for discipline specific writing courses. Lexical bundles are extremely common in academic discourse, and researchers consider them to be an important linguistic competence of fluent and native-like production.

Chen and Baker (2010), for example, investigated the structure and function of four-word lexical bundles in writings by Chinese students, native students, and native expert writers. They found that certain lexical bundle structures, such as NP-based lexical bundles (e.g., *in the context of*) were underused in both student essays. However, these two sets of students' essays overused the lexical bundles (e.g., *all over the world*). In all, the students' essays can be characterized as immature when overusing VP-based bundles whereas expert essays exhibit a wide range of NP-based bundles. Regarding the lexical bundle function, the researchers reported that non-native Chinese English writers do not rely on referential expression lexical bundles (e.g., *in the context of*, *the extent to which*, *at the same time*), which are frequently used in expert academic writing. In the comparison of writing proficiency, regardless of the L1 and L2, both the British university students and Chinese university students used a relatively higher number of discourse organizers compared to the professional academic writers.

Of the many purposes of lexical bundle studies, the pedagogical purposes have been of much interest to language researchers and practitioners because corpus-based linguistic analysis enables researchers and practitioners to study L2 learners' systematic linguistic characteristics. Granger (1998) explained that corpus linguistics focuses on language performance, specific description, and quantitative analysis enabling empirical research in second language acquisition (SLA) to reveal the governing processes of language learning. Thus, L2 corpora have been examined not to pinpoint the L2 learners' errors, but to understand how their language develops (see Meunier, De Cock, Gilquin, & Paquot, 2011, for an extensive discussion about L1 and L2 corpora research).

The discussion of the overuse and underuse of lexical items in L2 language might suggest crucial implications for L2 academic writing education. In the earlier lexical uses research, Ringbom (1998) compared the frequency of lexical uses in English native speakers' argumentative essays and learners' essays written by non-native speakers of English. The results showed that language learners either underused or overused certain vocabulary, and the non-native speakers' writings showed that L2 essays contained a smaller range of vocabulary than L1 essays do. In addition, the researcher concluded that certain lexical items are overused/circulated in the L2 in the learner corpora.

In their L2 corpus study, Granger, Dagneaux, Meunier, and Paquot (2009) provided 16 L2 argumentative writing corpora according to L1 backgrounds including not only Western languages (e.g., Czech, Dutch, French, and Russian), but also Asian languages (e.g., Chinese and Japanese). The L2 corpora sets allowed for comparisons between L1 and L2 and between one L2 and another L2. The comparisons led to the explanation about the characteristics of L2 writing. In the same manner, the study of Korean L2 corpus would make valuable contributions to understanding and characterizing learners' lexical bundle use and developing learner-specific pedagogical implications.

L2 learners' use of lexical bundle study is important in analyzing L2 writing and developing pedagogical implications of English academic writing for L2 learners, because it reveals how L2 learners use the recurrent expressions in academic writing as crucial criteria for demonstrating their writing proficiency with disciplinary conventions in the academic genre. Although research so far has suggested a structural discrepancy of lexical bundles between L2 writing and native speakers' writing (Chen & Baker, 2010; Öztürk & Köse, 2016), it is still vital to broaden our understanding of an alternative taxonomy of the lexical bundles as a long list of lexical bundle structures may not suggest pedagogical implications

(Güngör & Uysal, 2016). Furthermore, the underuse, overuse, and misuse of lexical bundles are outstanding characteristics of L2 writers, which may hinder L2 writers' academic writing proficiency. Although the studies of lexical bundles have explored diverse perspectives in L1 and L2 writing, the contrastive analysis lexical bundle patterns in L1 and L2 may uncover the underuse and overuse of the lexical bundle usages (see Granger et al., 2009, for a corpus-based contrastive analysis).

Drawing upon a crucial role of lexical bundles in L2 writing and their theoretical and practical potentials in language education, the current study aims to compare the functional distribution of lexical bundles in academic writing across two language speakers: English L2 speakers and English native speakers. In this regard, the following question is addressed in this study: To what extent do Korean and native English college students' differ in terms of type and token frequency of the lexical bundles and their functions?

Method

Data

Two kinds of corpora were prepared in the present study: a corpus of systematically collected Korean college students' argumentative essays, KRUNIV-Arg corpus, and its counterpart, MICUSP-Arg, which was compiled with the argumentative essays collected from the Michigan Corpus of Upper-Level Student Papers, the MICUSP (Römer & Wulff, 2010), a high-proficiency academic English corpus compared to L2 English writing. For KRUNIV-Arg, argumentative essays were systematically collected from four different universities in South Korea. The contributors of the essays were juniors and seniors of English language and literature, English education, and their related disciplines, such as linguistics. To ensure a fair comparison between the corpora, MICUSP-Arg was customized and restructured from the sub-corpora of MICUSP. For example, MICUSP includes a total of 829 essays consisting of 1.6 million words across 16 disciplines, including humanities, social sciences, natural science, and engineering. Because the essays in KRUNIV-Arg are argumentative writing from the specific disciplines, the argumentative essays in MICUSP from the three disciplines of English, education, and linguistics were selected and recompiled. In addition, since the corpus includes non-native English speakers' contribution and graduate students' writing, these essays were excluded to guarantee the fair comparison between the argumentative essay corpora. Table 1 presents the overview of the two argumentative essay corpora.

TABLE 1
Constituents of the Argumentative Writing Corpora

Corpus	Representation	Word count	Average text length	Number of texts
KRUNIV-Arg	English learner argumentative writing	201,241	564.4	358
MICUSP-Arg	Upper-level argumentative writing	106,659	2,091.4	51

In the corpora, the size, the average length, and the number of essays are different. However, since the corpora consist of argumentative essays prepared in similar academic disciplines, controlled and normalized size of the corpora is expected to ensure the validity of the current corpus analysis. This procedure is described in the following sections.

Categorization of Lexical Bundle Functions

The functional categorization in the present study follows the lexical bundle taxonomy proposed by

Biber et al. (2004) and Chen and Baker (2010). Based on the functional taxonomy of the lexical bundles, an experienced college English instructor and the researcher independently categorized the extracted lexical bundles. Although there was little disagreement between the raters, when there was disagreement over determining the categorization, the two experts discussed the determination of the function to come to an agreement. Three major categories of the lexical bundles are stance bundles, discourse organizers, and referential expressions. In each major category, several sub-categories can be found. In addition, the lexical bundles that do not fit in these categories are saved in the "others" category.

Stance lexical bundles serve a frame for the preposition of delivering two kinds of meanings, such as epistemic and attitudinal/modality functions: epistemic; desire; obligation/directiveness; intention/prediction; and ability:

- Epistemic: it is true that, I think that the, I don't agree with
- Desire: if you want to, people want to go, do what they want
- Obligation/directiveness: the most important thing, we don't have to, we should think that
- Intention/prediction: I would like to, it will be, to be effective in
- Ability: not be able to, it is impossible to, hard to get a

Discourse organizers include the lexical bundle functions of topic introduction/focus, topic elaboration/classification, and identification/focus function lexical bundles:

- Topic introduction: first of all, I want to say, there has been a
- Topic elaboration: on the other hand, that is to say, for example there are
- Identification/focusing: is one of the, is the best way, is no longer a

Referential expression lexical bundles include specification attributes, imprecision, and time/place/text deictic references:

- Imprecision: some people argue that, some might argue that, some are more equal
- Specification attributes
- Quantification: has a lot of, for the rest of, most of them are
- Tangible/intangible framing attribute: as a result of, this point of view, the effect of the
- Time/place/text-deictic (reference): for a long time, when I was a, from the beginning of

Finally, lexical bundle functions that do not fit in the categories above or ones with special conversational functions are categorized as "others": *he or she is, what do you think.*

Procedure

In corpus-based studies, frequency should be carefully handled to prevent (or at least minimize) possible skewed data analysis due to the comparison of corpora of different sizes. In fact, MICUSP-Arg is about 47% smaller than KRUNIV-Arg. According to Chen and Baker (2010), a normalized frequency of lexical bundle cut-off arbitrarily ranges between 20 and 40 times per million words, although for a smaller corpora such as a spoken corpus, raw cut-off ranges from two to 10 are considered (De Cock, 1998). Regarding the number of different texts where certain lexical bundles occur, previous studies have suggested three to five texts (e.g., Biber & Barbieri, 2007) or 10% of the whole text (e.g., Hyland, 2008a). The last issue concerning selecting lexical bundles is the length of the lexical bundles, usually ranging from two- to six-word combinations. Of the combinations, four-word lexical bundles are considered to be the most researchable bundles because the number of four-bundles can be within a manageable size for manual categorization and concordance observations (Chen & Baker, 2010).

After conducting a set of pilot analyses, the cut-off point for the frequency of the four-word lexical

bundles was set to five times or more for KRUNIV-Arg and three times or more for MICUSP-Arg, occurring in at least three texts. In fact, as discussed in the studies already mentioned, the number of frequency cut-off points should be determined based on sufficient representatives of the corpora being investigated. One of the techniques for the fair comparison of the lexical bundles from different corpora is to identify a standardized threshold or normalization. In the current study, the cut-off point was set based on 25 times per million words, which allowed a fair comparison between the two corpora of different sizes. Table 2 shows the relationship between the normalized and the raw frequency of the thresholds. The decimal numbers in the normalized frequency rounded up for the operational minimum cut-off frequency.

TABLE 2
Normalized and Raw Frequency Thresholds

Corpus	Set normalized frequency threshold (per million)	Corresponding raw frequency threshold (per million)
KRUNIV-Arg	25	5.0
MICUSP-Arg	25	2.6

The concordance program *AntConc 3.4.4* (Anthony, 2014) was used to generate lexical bundles automatically. In corpus-based studies, frequency data should be carefully handled to avoid any skewed data due to the comparison of corpora of different sizes. A process of frequency normalization was employed to guard against any possible data distortion.

A series precautionary measure was taken into consideration qualitatively and quantitatively: Word sequences pertaining to the argumentative essay questions (e.g., *a college degree necessary*, *a foreign language as*) or context-dependent bundles, usually with proper nouns (e.g., *The Vicar of Wakefield*, *Cambridge Harvard University Press*), were manually checked and excluded from the extracted bundle lists. If these bundles were included in the corpora, they would inflate the results of the quantitative comparison analysis. There is another point to consider when counting the number of bundles. For instance, two sequences of words may be a part of a longer sequence of words. In the current research, one corpus included five lexical bundles of *important to note that* and *is important to note*. The two sets of the bundle cannot be considered two different bundles as they are part of a longer lexical bundle, *is important to note that*. Therefore, these kinds of overlapping bundles are counted as they occur (i.e., five times). Another example of the bundle requiring adjustment is the bundles like *lot of people who* and *a lot of people who* occurring five times each. In fact, the word sequences of these bundles are included in *a lot of people*, which occurs 18 times. Therefore, the lexical bundle *lot of people who* and similar types of lexical bundles were excluded in the counts so as not to inflate the results. Table 3 describes the refined number of lexical bundles after the adjustments.

TABLE 3
Number of Bundles Before and After the Refinement

Corpus	Before refinement		After refinement	
	Number of lexical bundles (types)	Number of lexical bundles (tokens)	Number of lexical bundles (types)	Number of lexical bundles (tokens)
KRUNIV-Arg	466	3,898	231	1,820
MICUSP-Arg	116	615	89	464

Results

The purpose of this study was to identify the different use of lexical bundles by comparing and contrasting their functional distributions in Korean and American college students' argumentative writing corpora. To this end, after the automatic extraction and the refinement process of the lexical bundles in

the corpora, the lexical bundles from the Korean and American university students were categorized according to their functions. The number of the lexical bundle types and tokens were then compared to examine the distributional differences and similarities in each writing. For this comparison, the proportions of the lexical bundle function categories were compared for both types and tokens, then chi-square tests were conducted to confirm how the functional distribution of the lexical bundles can be characterized. The standardized residuals (*R*) were also examined to determine which lexical bundle functions are major contributors to the lexical bundle usage difference between KRUNIV-Arg and MICUSP-Arg.

Lexical Bundle Function Types in Argumentative Writings

As shown in Figure 1, MICUSP-Arg contains a higher proportion of referential expressions (55%), while the expressions are much less frequent in KRUNIV-Arg (26%). On the other hand, stance lexical bundles (35%) and discourse organizers (33%) in KRUNIV-Arg occur more frequently than in MICUSP-Arg.

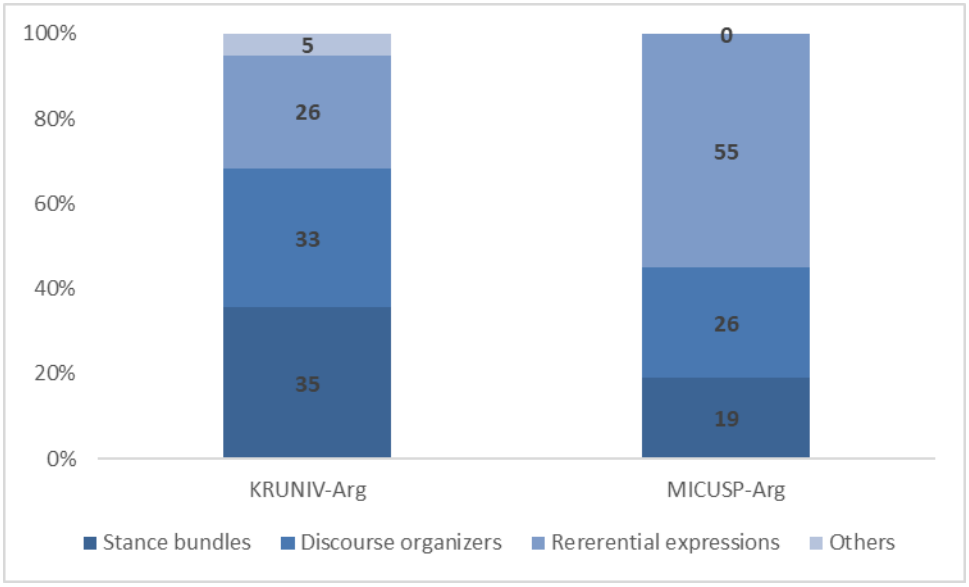

Figure 1. Functional distribution of lexical bundle types

The chi-square test indicated a significant difference regarding the functional distribution of the lexical bundle types (χ^2 = 26.581; *df* = 3; *p* = 0.000; Cramer's V = 0.288). The standardized residuals were also calculated to determine which lexical bundle function made a major distribution to the difference.

TABLE 4
Chi-Square Contingency Table for Functional Distribution: Types

		Stance bundle	Discourse organizers	Referential expressions	Others
KRUNIV-Arg	Observed count	82	76	61	12
	Expected count	71.47	71.47	79.41	8.66
	R	1.25	0.54	-2.07	1.13
MICUSP-Arg	Observed count	17	23	49	0
	Expected count	27.53	27.53	30.59	3.34
	R	-2.01	-0.86	3.33	-1.83

Note. χ^2 = 26.581; *df* = 3; *p* = 0.000; Cramer's V = 0.288; Effect size = 0.29 (small effect size); Other lexical bundles

include special conversational bundles.

As presented in Table 4, the referential expressions in KRUNIV-Arg and MICUSP-Arg have an absolute value of *R* greater than 1.96. The absolute value of residuals greater than 1.96 is considered large (i.e., factors contributing to the difference). A large standardized residual value means that the observed frequency in the cell differs significantly from its expected frequency. The sign of the standardized residual indicates whether the observed frequency is above or below (numbers with the minus [-] sign) the expected frequency. Therefore, the results in Table 4 confirm that the referential expressions lexical bundles made a statistically significant contribution to the hypothesis that the KRUNIV-Arg and MICUSP-Arg are statistically different regarding the functions.

The results suggest that American college students use referential lexical bundles to make direct reference to physical/abstract entities or to the context to opt out of some particular attribute of the entity (e.g., *in the realm of, at the very least*). To the contrary, the Korean college students pay more attention to stance bundles for their evaluation or judgment (*I do not think, believe that it is*) and discourse organizing lexical bundles for structuring texts (*first of all we, that is why I*).

Lexical Bundle Function Tokens in Argumentative Writing

The token distribution of the functions between the corpora is more or less the same as for the type distribution. Figure 2 illustrates the proportion of the four functional categories according to the lexical bundle tokens. In MICUSP-Arg, the most frequent lexical bundle function is referential expressions (60%); however, in KRUNIV-Arg, discourse organizers (36%) account for the greatest number of the lexical bundle tokens, followed by stance lexical bundles (33%) and referential expressions (26%).

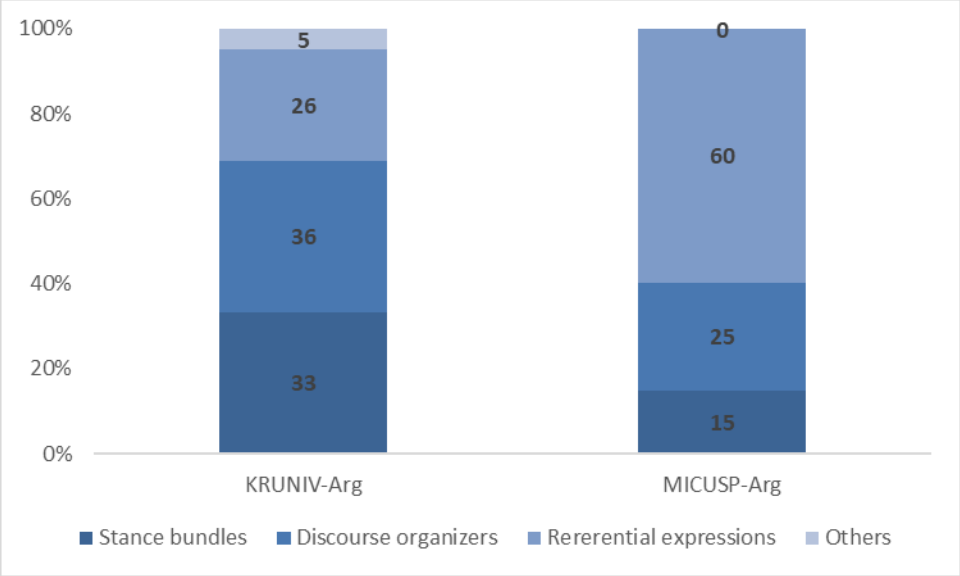

Figure 2. Functional distribution of lexical bundle tokens

A chi-square test confirms that a significant difference exists regarding the functional distribution of the lexical bundle between KRUNIV-Arg and MICUSP-Arg at the 0.001 level (χ^2 = 145.070; *df* = 3; *p*=0.000; Cramer's V = 0.255). The standardized residuals calculated also identified the major contributor(s) to the difference. Table 5 shows the stance bundle and referential expressions contributed to the difference. The result is slightly different from the standardized residual results of the lexical bundle type differences. After the lexical bundle token analysis, the two college student groups' use of the functional lexical bundle was more contrasting: Two of the lexical bundle categories are contributing factors, making the

two essay corpora significantly different. In the KRUNIV-Arg, stance bundle is more frequently used. On the other hand, referential expressions are more frequently used in MICUSP-Arg.

TABLE 5
Chi-Square Contingency Table for Functional Distribution: Tokens

		Stance bundle	Discourse organizers	Referential expressions	Others
KRUNIV-Arg	Observed count	607	648	475	90
	Expected count	550.73	624.05	571.91	70.32
	R	2.40	0.96	-4.05	1.95
MICUSP-Arg	Observed count	69	118	227	0
	Expected count	125.27	141.95	230.09	26.68
	R	-5.03	-2.01	8.50	-4.08

Note. χ^2 = 145.070; *df* = 3; *p* = 0.000; Cramer's V = 0.255; Effect size = 0.25 (small effect size); Other lexical bundles include special conversational bundles.

Although the standardized residual values of the discourse organizers and others are marginal, Table 5 clarifies the usage difference of the lexical bundles between the corpora. The contrast is probably due to the significant difference in the number of lexical bundle functional tokens occurring in KRUNIV-Arg and MICUSP-Arg. The comparison of Tables 4 and 5 suggests an interesting point for understanding how Korean college students use the lexical bundles in the argumentative writing. As revealed in the functional distribution of the lexical bundle types (Figure 1 and Table 1) and the token (Figure 2 and Table 2), in KRUNIV-Arg and MICUSP-Arg, referential expressions occur substantially frequently in the American college students' argumentative writing. On the other hand, in the Korean college students' argumentative writing, stance bundle is frequently used, showing a statistical significance.

Although the residual value is marginal, one noteworthy finding revealed in the lexical bundle token analysis is that the Korean college students' writing contains a considerable amount of conversational lexical bundles (*have to worry about*, *we are living in*), none of which is found in the American college students' writing. Thus, Korean college students may need to be aware of the formality of academic writing and practice it in their actual writing. Another interesting finding from the analyses lies in the lexical bundle type and token occurrence differences. The type/token ratios of the lexical bundle categories in KRUNIV-Arg are smaller than those of MICUSP-Arg, suggesting that Korean college students use a relatively narrow range of functional lexical bundles, reflecting the fact that the use of the lexical bundles is repetitive and simplistic. On the other hand, in MICUSP-Arg, the use of the lexical bundles is relatively unique and diverse.

Discussion and Conclusion

The current study of L1 and L2 lexical bundles uncovered valuable findings for understanding the characteristics of the Korean college students' English argumentative writing. The functional distribution comparisons of the lexical bundles showed that the use of the lexical bundles in Korean and American students' argumentative writing is significantly different. The functional distribution analysis of lexical bundle types found that, in American university students' writing, more than half of the lexical bundle types are referential expressions. The extensive use of referential lexical bundles in the American students' writing reflects how lexical bundles are used in written academic register. Of the functional lexical bundles, according to Biber and Barbieri (2007), the referential lexical bundle types are by far the most frequently used in textbooks and academic prose. On the other hand, in Korean university students' argumentative writing, the referential lexical bundles account for only a quarter of their functional lexical bundle use. In addition, the chi-test confirmed that the functional distributions of the lexical bundle types between the two corpora were significantly different with the major contributions of the referential

expressions. Relatively higher use of stance and discourse organizers is a common practice in novice academic writing. The findings of the current study also suggested the different use of functional lexical bundles in different proficiency levels. For example, Öztürk and Köse (2016) conducted a comparative lexical bundle study in non-native postgraduate, native postgraduate, and native scholars' writing and found that, even if the distribution of referential lexical bundles expands as the writers' proficiency levels increase, the distribution of stance and discourse organizer gets narrower. Their study also revealed that the proportion of stance and discourse organizer lexical bundles are similar. Therefore, the Korean college students' argumentative writing needs to be improved in terms of the use of referential lexical bundles.

For the functional lexical bundle token distributions, a similar pattern was found. In the American students' argumentative writing, more than 60% of the lexical bundles covered referential expressions. In Korean students' writing, the three functional lexical bundles (i.e., stance, discourse organizers, and referential expressions) covered about 95% of the functional lexical bundle uses. Another chi-square test confirmed that the lexical bundle token distributions were significantly different between the two argumentative writings. In addition to the referential lexical bundles, stance bundles significantly contributed to the difference. Given the relationship between the number of lexical bundle types and tokens in the functional categories, Korean college students' argumentative writing was found to be immature, having less referential lexical bundles. Mature/L1 writers probably use thematic structures of theme and rheme more proficiently in academic writing than novice/L2 writers, as Nam and Park (2015) concluded that the EFL writers use repetitive theme functions such as *first of all*, *second*, *moreover*, *as a result*, and *in conclusion*, which share similar functions of lexical bundles and discourse organizers. This finding can be partly corroborated by Chen and Baker (2010). According to their report on the lexical bundle functional distribution comparisons, novice academic prose writers—whether native or non-native—use a relatively higher number of discourse organizers; professional academic writers, on the other hand, use referential expressions significantly more than novice writers.

In the same vein, one of the major discoveries of comparative interlanguage analysis is the underuse/overuse of certain lexical items in L2 language (Granger, 2002; Ringbom, 1998). The present study of lexical bundle distributions found that the Korean college students overused the stance lexical bundles and discourse organizers. Regarding underuse, they less frequently used referential expressions in their academic writing than the upper-level students. In terms of the overuse of certain lexical items, Nam (2013) found that, in university-level argumentative writing, Korean learners of English use the personal pronoun *I*. The finding confirms why the stance lexical bundles (e.g., *I think that, I don't agree with, I strongly believe that*) and discourse organizers (e.g., *I want to say, in this essay I, for the reason I*) are more frequently used than in the American students' argumentative writing.

The results of the current study can be used to suggest pedagogical practices with the lexical bundles in academic writing for non-native English learners. Specifically, functionally related lexical bundles collected from authentic academic texts can be systematically and abundantly introduced to students. Given that Korean learners of English may not understand how upper-level students effectively use functional lexical bundles, explicit instruction on lexical bundle frequency and distribution would be useful for raising learners' awareness.

Language researchers and practitioners have been interested in the investigation of the lexical bundles in academic writing. Although there has been much research on the analysis of writing products, there is little research on the pedagogical application of the lexical bundles. Of the small number of studies, Cortes (2006) conducted a study focused on the teaching of a group of lexical bundles in a writing-intensive university history class. Although the study found that there is no difference between the pre- and post-instruction of lexical bundle production, the findings indicated an increase of students' awareness and interest in the lexical bundles. Once learners are aware of the frequent and appropriate way of incorporating lexical bundles in their writing, it is also important for them to analyze the use of the recurrent word combinations in different academic settings, such as in novice and professional writing or in different disciplines. Certain academic communities' awareness of lexical bundles could help learners uncover a new way of helping themselves become active community members.

The lexical bundles consist of structure and function constituents as a unit of language. The present research found and confirmed the important characteristics of Korean learners' English writing towards the use of lexical bundle functions. The conclusion, however, does not answer the directionality of the lexical bundle uses—namely, whether the frequent use of the personal pronoun *I* leads to proportional uses of stance lexical bundles and discourse organizers or vice versa. The question might be answered with sociolinguistic or cultural linguistic investigations in future studies.

The Author

Daehyeon Nam is Assistant Professor in the Division of General Studies at Ulsan National Institute of Science and Technology (UNIST) in S. Korea. His current research focuses on EAP/ESP writing pedagogy though corpus linguistics and its application to genre and discourse analysis. His recent publications include *Lexical bundle structures of nuclear science and engineering research article* (2017), *Corpus-based analysis of thematic structure in L2 writing* (2015; co-authored). Currently, he is conducting research on the citation/network analysis of corpus linguistics research articles and the corpus-based Korean news interview analysis.

Division of General Studies
Ulsan National Institute of Science and Technology
50 UNIST-gil, Ulju-gun
Ulsan, 44919, Republic of Korea
Tel: +82-52-217-2023
Email: dnam@unist.ac.kr

References

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

Biber, D. (2006a). Stance in spoken and written university registers. *Journal of English for Academic Purposes, 5*(2), 97–116.

Biber, D. (2006b). *University language: A corpus-based study of spoken and written registers.* Amsterdam, the Netherlands: John Benjamins.

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

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

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

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing in the teaching of EAP. *University of Sydney Papers in TESOL, 5*(5), 31–64.

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

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397–423.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education, 17*(4), 391–406.

Cowie, A. P. (1998). *Phraseology: Theory, analysis, and applications*. Oxford, UK: Oxford University Press.

De Cock, S. (1998). A recurrent word combination approach to the study of formulaic in the speech of

native and non-native speakers of English. *International Journal of Corpus Linguistics, 3*(1), 59–80.

Durrant, P. (2015). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics, 36*(1), 1–30.

Firth, J. R. (1957). *Papers in linguistics, 1934–1951*. London, UK: Oxford University Press.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam, the Netherlands: John Benjamins.

Granger, S. (Ed.). (1998). *Learner English on computer*. London, UK/New York, NY: Longman.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.). (2009). *International corpus of learner English: Version 2*. Leuven, Belgium: Presses universitaires de Louvain.

Güngör, F., & Uysal, H. H. (2016). A comparative analysis of lexical bundles used by native and non-native scholars. *English Language Teaching, 9*(6), 176–188.

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

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

Manning, C., & Schütze, H. (2001). *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

Meunier, F., De Cock, S., Gilquin, G., & Paquot, M. (Eds.). (2011). *A taste for corpora: In honour of Sylviane Granger*. New York, NY/Amsterdam, the Netherlands: John Benjamins.

Nam, D. (2013). A corpus-based contrastive analysis of Korean college students' English composition. *English Language Teaching, 25*(4), 67–85.

Nam, D., & Park, K. (2015). Corpus-based analysis of thematic structure in L2 writing. *Multimedia-Assisted Language Learning, 18*(4), 99–120.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford, UK: Oxford University Press.

Öztürk, Y., & Köse, G. (2016). Turkish and native English academic writers' use of lexical bundles. *Journal of Language and Linguistic Studies, 12*(1), 149–165.

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.

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.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–225). London, UK: Longman.

Reppen, R. (2009). Exploring L1 and L2 writing development through collocations: A corpus based look. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language* (pp. 49–59). Basingstoke, UK: Palgrave Macmillan.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on computer* (pp. 41–52). London, UK/New York, NY: Longman.

Römer, U. (2010). Using general and specialized corpora in English language teaching: Past, present and future. In M. C. Campoy-Cubillo, B. Bellés-Fortuño, & M. L. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching* (pp. 18–38). London, UK: Continuum.

Römer, U., & Wulff, S. (2010). Applying corpus methods to written academic texts: Explorations of MICUSP. *Journal of Writing Research, 2*(2), 99–127.

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing*. New York, NY/Amsterdam, the Netherlands k: John Benjamins.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. New York, NY/Amsterdam, the Netherlands: John Benjamins.

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

Wei, Y., & Lei, L. (2011). Lexical bundles in the academic writing of advanced Chinese EFL learners. *RELC Journal, 42*(2), 155–165.

Wray, A. (2006). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press.