Modern English Education, Vol. 16, No. 3, Fall 2015
http://dx.doi.org/10.18095/meeso.2015.16.3.03

Lexical Bundles in Korean University Students' EFL Compositions: A Comparative Study of Register and Use

Choongil Yoon[*]
Ewha Womans University
Ji-Myoung Choi
Yonsei University

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

Lexical bundles are considered important building blocks of discourse as they perform a wide range of discourse functions serving as markers of fluency and appropriacy of register-specific language use (Biber, 2009; Hyland, 2012). To investigate the extent to which this is the case with Korean university students' English essays, the present study identified and analyzed four-word lexical bundles from a corpus of EFL argumentative essays written by Korean university students and compared the bundles identified with those from a corpus of native speaker (NS) students' argumentative writing. Analysis of the bundles across grammatical structures and discourse functions revealed that compared to NS students, the Korean university students heavily relied on the bundles widely used in speech registers such as bundles containing personal pronouns and contractions and expressing stance. On the other hand, the Korean students' essays were shown to have fewer incidences of nominalizations, hedging and referential bundles. A concordance analysis of *on the other hand*, the most frequently used bundle, showed a substantial number of erroneous and ineffective uses of the bundle as well. The paper concludes by discussing pedagogical implications for EFL composition pedagogy and suggestions for future studies.

**[lexical bundle/register/EFL composition/
어휘다발/사용역/영어작문]**

[*] First author: Choongil Yoon, Corresponding author: Ji-Myoung Choi

I. INTRODUCTION

It is now widely recognized that formulaic language—an umbrella term covering a whole range of multi-word units including idioms, collocation, phrase verbs, and other (semi-)fixed expressions—comprises a large portion of normal discourse, performing a wide range of functions (Altenberg, 1998; Schmitt, 2010) and that it contributes to fluent and native-like language processing and use (Paquot & Granger, 2012). While different methods have been employed to identify and analyze different types and aspects of formulaic language (Schmitt, 2010), one of the most frequently used methods is corpus-based investigation of lexical bundles. Lexical bundles are multi-word sequences that occur at a certain frequency across a range of texts (Wood, 2010), and thus represent the most frequently recurring word sequences in a collection of texts. This frequency-based corpus method has an advantage of being methodologically clear-cut compared to phraseological approaches (Hyland, 2012), which often require extensive analysis of relatively small amount of data using human judgement. Coupled with this methodological advantage, the increasing availability of corpora varied in type and size and user-friendly corpus tools have led to a wealth of studies in lexical bundles in recent years.

These studies have revealed that lexical bundles, extracted in the form of continuous word sequences with a fixed number of words (or n-gram) are usually neither idiomatic nor psychologically salient, unlike other types of formulaic language (Biber, 2009), but that they are important building blocks in discourse performing various communicative and discourse functions and thus act as markers of proficiency in a given genre and register (Hyland, 2012; Wood, 2010). Lexical bundle studies have often taken a contrastive approach comparing different registers, and different levels of language proficiencies. When it comes to L2 learning and teaching in particular, many studies have compared lexical bundles retrieved from learner corpora with those of native speakers', offering insights into common and unique patterns of learner phraseologies across different L1s. Findings of these studies suggest that L2 learners may greatly benefit from raising awareness of phraseological patterns needed for fluent, functional, and communicative language use.

The present study aims to add to this growing body of learner lexical bundle studies by conducting a systematic analysis of lexical bundles from a corpus of Korean university students' EFL compositions. Particularly, the study focuses on the frequencies of the lexical bundles, their functional and structural patterns, and the register characteristics drawn from those patterns, and explores the Korean students' overuse, underuse, and misuse of lexical bundles in comparisons with native speaker (NS) students.

II. LITERATURE REVIEW

1. Identification and Classification of Lexical Bundles

Most previous research follows the frameworks established by Altenberg (1998) and Biber, Johansson, Leech, Conrad, and Finegan (1999) in their seminal studies that identified frequency-based recurrent word sequences and analyzed them in terms of grammatical structures and discourse functions. To begin with, lexical bundles, by definition, are identified commonly by two criteria: how recurrent a bundle should be (i.e., a minimum frequency) and how widely it should be used (i.e., the number of texts or files in which it appears). There has been no consensus on the threshold numbers for these criteria as different studies employed a different set of frequency and range cutoff points: 10 to 40 times per million words for frequency and three to five texts for range. However, some researchers (Chen & Baker, 2010; Hyland, 2012) caution against working with a small corpus or comparing bundles obtained from corpora of widely different sizes as small corpora tend to produce greater normalized numbers of bundles. Another important number in a lexical bundle study is the length of word strings to be analyzed, usually presented as n-gram. Different lengths ranging from bigram to six-gram have usually been investigated but four-gram sequences have been the most frequently used as they are considered to produce a sufficient variety of structures and functions to analyze while still being manageable in terms of data amount to deal with.

Once lexical bundles are identified with a set of these pre-determined numbers, they are mostly classified according to their grammatical structures and discourse functions. Most studies employed the taxonomies developed in Biber et al. (1999) or their slightly modified versions with subcategories added or merged. For structure classification, bundles have commonly been classified into noun phrase (NP)-based, prepositional phrase (PP)-based and verb phrase (VP)-based bundles with each having multiple subcategories. For discourse functions, most of previous studies of written data classified their lexical bundles into three major categories. Firstly, referential bundles are those expressions making reference to entities or describing their attributes. Secondly, stance bundles express the writer's knowledge of or attitude toward the proposition that follows or assessment of its certainty. Lastly, discourse organizers are used to define relationships with preceding or following discourse in text performing specific functions such as topic introduction and elaboration. It should be noted, however, that some categories of discourse functions and their subcategories used in these studies have not been clearly defined inviting much arbitrariness or subjectivity in classification (Ädel & Erman, 2012).

One final procedure that has commonly been carried out in many studies is exclusion of "prompt/content bundles" (e.g., *in the United States*, *agree with the following statement*)

that are closely related to essay prompts or specific to essay topics (Chen & Baker, 2010; Staples, Egbert, Biber, & McClair, 2013). This procedure is considered necessary because such bundles are simply products of the specific topics and prompt questions of the essays that offer little about learners' general recurrent language use patterns. Some studies also excluded overlapping sequences that are both part of the same longer bundles (Ädel & Erman, 2012; Chen & Baker, 2010), which otherwise would inflate the number of bundles.

2. Differences across Registers and Different Groups of Language Users

Starting from their work for the *Longman Grammar of Spoken and Written English* (Biber et al., 1999), Biber and his colleagues have explored register variations in lexical bundles across spoken and written registers (and their subtypes) through a series of studies (Biber, 2009; Biber & Barbieri, 2007; Biber, Conrad & Cortes, 2004). These studies found that while lexical bundles are overall more frequent and varied in spoken than written registers, the two registers are dominated by different structural and functional bundles. Spoken data were shown to contain large portions of clausal phrases and VPs in terms of structure and stance bundles and discourse organizers in functions whereas written prose was predominated by NPs, and PPs and by referential bundles. These findings have served as an important analytic framework in other subsequent lexical bundle studies.

Drawing on the frameworks developed by Biber and colleagues, many studies have examined variations in the use and distribution of lexical bundles across different groups of language users such as 1) native versus non-native speakers, 2) novice versus expert writers, and 3) language learners at different proficiency levels. Due to the far greater availability, most of these studies were done with corpora of written data, especially academic prose. First, there are a considerable number of studies that compared lexical bundles extracted from a learner corpus and a NS corpus (mostly English) (e.g., Ädel & Erman, 2012; Granger, 1998; Chen & Baker, 2010) identifying bundles over- and under-used by learners. In terms of frequency of lexical bundles, results were mixed. While some studies revealed that learners used more lexical bundles than native speakers, often involving repeated use of a small set of bundles (S. C. Hong, 2013, Paquot & Granger, 2012), in other studies (Ädel & Erman, 2012; Chen & Baker, 2010) learners' use of bundles was less frequent and less varied. While pointing out a tendency of less proficient learners to copy phrases directly from essay prompts, Paquot and Granger (2012) suggested that the frequency and distribution of lexical bundles can be influenced by the situational factors such as the topics and prompt questions, and the writer's language proficiency.

In terms of the specific bundles over- and under-used by L2 writers of English, studies commonly found the dominance of lexical bundles typical of speech. In a study examining

EFL argumentative essays written by students of different L1 backgrounds, Granger (1998) reported, for example, L2 writers' tendency to overuse *I/we* framed constructions but to underuse conventionalized adverb + adjective combinations (e.g., *painfully clear, readily available*). Chen and Baker (2010) found that in comparison with native speaker students and published authors, Chinese ESL writers used no or fewer bundles that can be seen as hedging devices (e.g., *it is likely to, would have to be,* and *to a large extent*) and frames like "*the* Noun *of the*", and "*in the* Noun *of*" that are typically frequent in academic prose, while using bundles that can be taken to overgeneralize (e.g., *all over the world*) and be unnecessarily verbose (e.g., *that is to say*). Ädel and Erman (2012) also found that Swedish L1 students used less hedges and passives in their EFL argumentative essays than their NS counterparts.

There are a small number of studies that examined lexical bundle variances among populations of different language/writing proficiencies. Cortes (2004) compared English L1 students' writing with published journal articles and found that the students rarely used the academic lexical bundles identified from the journal articles but when they did, they often used the bundles inappropriately. Staples et al. (2013), on the other hand, investigated the frequency and functions of lexical bundles in written responses in the TOEFL *i*BT test across three different proficiency levels. The results showed that lower level students used more bundles, many of which came from the prompts. The results of these studies suggest that the register confusion (i.e., using lexical bundles inappropriate to the register of the text) is as much a developmental issue that affects novice writers as it is part of learning a foreign/second language.

3. Lexical Bundles from Korean EFL Learner Corpora

There have been only a few studies that directly investigated the patterns of lexical bundles from Korean EFL learners' data. Lee (2009) and Kwon and Lee (2014) analyzed lexical bundles extracted from a spoken corpus of teaching demonstrations by Korean university students and Korean EFL teacher talk respectively. By comparing and contrasting against corresponding spoken corpora of native speakers, the two studies commonly showed that the Korean learners overused limited sets of bundles that were mostly clausal in structure and stance related in function.

J. H. Kim (2013) and S. C. Hong (2013, 2015), on the other hand, compared lexical bundles from a corpus of Korean university students' English argumentative writing with those from a corresponding NS students' corpus. The Korean student writers were shown to overuse the "personal pronoun + verb" frame mostly expressing one's stance (e.g., *I think it is, I do not agree*) but underuse NP and PP bundles, which is consistent with the findings from other studies on learner bundles.

While these studies are significant in that they are, to my best knowledge, the first to look into lexical bundles from composition data of Korean EFL learners, however, the studies are, to some extent, limited in providing a clear snapshot of lexical bundles used in Korean university students' EFL argumentative writing. J. H. Kim's (2013) study, while conducting a solid analysis with the given data, the size of the corpus used is rather small (a little over 60,000 words), may not contain a sufficient range of potential bundles used by Korean EFL writers. S. C. Hong's (2013, 2015) studies, on the other hand, used a larger corpus that is comparable to the reference corpora used but seem to lack rigor in the identification and classification of bundles and in reporting on the results. For example, it is not clear whether prompt/content and overlapping bundles were excluded from analysis, a procedure that may have had a significant impact on the results. In addition, large percentages of learner bundles were classified into "Others," especially in discourse functions (40% - 60%), calling into question the point of using a taxonomy that cannot account for half of the data.

Motivated by these gaps, this study aims to apply a rigorous analytic framework in identifying and analyzing lexical bundles from a corpus of Korean university students' EFL argumentative writing, then in comparing and contrasting them with those retrieved from a corpus of native speaker students' argumentative writing. In so doing, the study goes beyond the simple raw frequency-based analysis 1) to explore statistically overused and underused bundles and 2) to investigate how some of the bundles are actually used in text. Specifically, it is guided by the following research questions: 1. What are the frequencies of lexical bundles in the corpus of Korean university students' EFL compositions and what are their patterns in terms of grammatical structure and discourse function? 2. What do those patterns identified tell us about the register of Korean university students' EFL argumentative writing? 3. What specific lexical bundles do Korean university students overuse, underuse, and, if any, mis-use in their writing, in comparison with NS university students?

III. METHOD

1. Corpus Data

The data used for the study were two corpora of argumentative writing drawn from two larger corpora: Neungyule Interlanguage Corpus of Korean Learners of English (NICKLE) and the Louvain Corpus of Native English Essays (LOCNESS). NICKLE is a million word corpus compiled as part of the Neungyule-Longman English-Korean Dictionary project, which was developed jointly by Neungyule Education Inc. of Korea and Pearson Longman

of the UK in 2009. It consists of seven different genres with 10% being spoken data. The source text data were collected from the first- and second-year university students at lower-intermediate to (upper-) intermediate levels. Unlike other Korean learner corpora used in previous studies, which were mostly compiled from data produced in a single institution, NICLKLE data were drawn from multiple universities across Korea with the help of the Korea Association of Teachers of English. For the present study, only the argumentative essay component was used for analysis. The argumentative essays included were mostly on popular and familiar topics such as euthanasia, abortion, and death penalty.

LOCNESS is a collection of English L1 students' writing, made up of literary/expository and argumentative essays from British and American university students. The corpus, built as part of the International Corpus of Learner English (ICLE) to serve as a reference NS corpus, is highly comparable to NICKLE in terms of text types, author profiles, and topics. For the study, only the argumentative essays were selected for analysis. The overview of data used in this research is presented in Table 1.

TABLE 1

The Two Corpora

Corpus	Word token	Word type	No. of Texts
NICKLE_argumentative*	193870	11219	429
LOCNESS_argumentative*	230190	12931	321

Note. * The two corpora are hereafter referred to as NICLKE and LOCNESS respectively for brevity.

2. Lexical Bundles Extraction and Refinement

To get results that are comparable to previous studies, we decided to analyze four-word bundles, which have been the most researched n-grams from L2 argumentative essays (Ädel & Erman, 2012; Chen & Baker, 2010; Cortes, 2004; Staples et al., 2013). For the first step, after several trials with different numbers, the cut-off points were set to five times for frequency (approximately 23 times per million words on average), occurring in at least four essays. Although the two corpora were different in size, we decided that the difference was not large enough to have different cut-off points for frequency (see Table 1). Applying these criteria, 378 lexical bundles were retrieved from NICKLE and 282 from LOCNESS. AntConc software was used for bundle extraction.

For the next step, as done in Chen and Baker (2010), and Ädel and Erman (2012), we removed topic-dependent bundles (i.e., prompt and content bundles) from the bundles extracted above. It turned out that a good number of bundles came straight from essay prompts (e.g., *own a cellular phone* from the topic "the advantages and disadvantages of owning a cellular phone", and *of the death penalty* from "death penalty"), or were closely related to the topics being discussed (e.g., *cars on the road* from essays about traffic

congestion) in both corpora.

Not to inflate the number of lexical bundles, additional data refinement was conducted. First, contractions (e.g., *don't have to, it's hard to*) and their non-contracted forms (e.g., *do not have to*, *it is hard to*) were counted as the same types (but represented by the more dominant form in the final bundle lists). Second, following the procedure done in Chen and Baker (2010), lexical bundles were manually checked for overlap and combined into one in the following two types of overlap: first, "complete overlap" where two four-word bundles are derived from a single five-word bundle (e.g. four-word bundles of *there are a lot* (8 tokens) and *are a lot of* (8) coming from the five-word bundle *there are a lot of*). Second, "complete subsumption" where the occurrences of one bundle subsume those of another bundle (e.g., the bundle *to the fact that* (18) subsumes the bundle *due to the fact* (16)). After the refinements, the numbers of bundle types were reduced to 188 for the Korean and 204 for the NS data.

3. Analysis

To analyze the characteristics and uses of the bundles identified and to compare them between the two corpora, the bundles were classified in terms of grammatical structures and discourse functions. For grammatical structures, we followed the taxonomy in the *Longman Grammar of Spoken and Written English* (LGSW), which classifies structures of lexical bundles into 12 categories in academic prose and 14 categories in conversation, which can, in turn, be broadly grouped into three major categories: NP-, VP-, and PP-based types. For this study, which aimed to examine the register characteristics of learner bundles, we combined the categories of the two registers above into 15 subcategories. As for discourse functions, we used the taxonomy in Biber et al. (2004) and its slightly modified versions (Biber & Barbierie, 2007; Chen & Baker, 2010) consisting of referential expressions, stance expressions, and discourse organizers (for descriptions of these categories, see Literature Review above). These three function groups were then classified into 13 subcategories according to the specific functions each bundle performs in text. Detailed descriptions of the two taxonomies and example bundles of each subcategory appear in Results and Discussion below when relevant findings are discussed (Tables 3, 4).

Finally, log likelihood (LL) was calculated for each bundle across the two corpora to identify bundles that occur unusually frequently (i.e., overused) or infrequently (i.e., underused) in NICKLE relative to LOCNESS. In previous studies (e.g., Y. E. Kwon & E. J. Lee, 2014), the Keywords function in WordSmith was used to retrieve overused and underused bundles, which is based on LL statistic. Since AntConc was used for this study, we calculated LL values using Excel.

IV. RESULTS AND DISCUSSION

Results are presented roughly in order of the research questions posed above, but the second research question regarding the register characteristics is discussed through the section wherever relevant.

1. Frequencies of Bundles

Four-gram extraction followed by a refinement procedure described in the Method section produced a total of 188 types and 1690 tokens of lexical bundles from NICKLE and 204 types and 1610 tokens from LOCNESS. Out of these bundle types, the two groups shared 53, just over one in every four bundle types from each corpus. This result indicates that while Korean students used a slightly greater number of bundles in terms of tokens, NS students showed a greater variety in their bundle use (see the type-token ratios in Table 2). Not as clear as in Ädel and Erman (2012) and Chen and Baker (2010), this pattern of greater variety in bundle types by native speakers has been confirmed in this study as well.

TABLE 2

Frequencies of Lexical Bundles from Two Corpora

Corpus	Type	Token	TTR
NICKLE	188	1690	0.11
LOCNESS	204	1610	0.13

2. Grammatical Structures

Structural classification was conducted following Biber et al. (1999). As described in Method, an integrated version of their two taxonomies (academic prose and conversation) was used. Figure 1 below represents the proportional distribution of bundles within each corpus by category and Table 3 shows the distribution of lexical bundles across subcategories of grammatical structure.

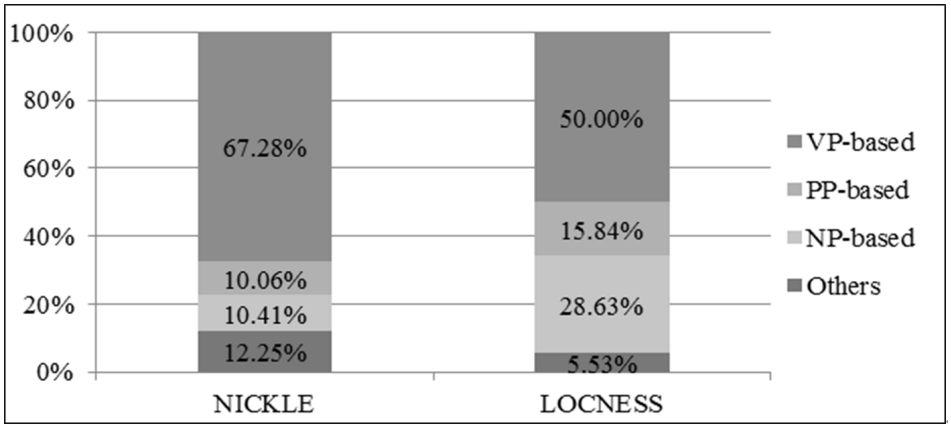

FIGURE 1 Distribution of Bundles Across Grammatical Structures in Each Corpus (token)

TABLE 3

Distribution of Lexical Bundles Across Structural Subcategories (token)

Category	Subcategory	NICKLE	LOCNESS	Example
NP-based	NP with *of*-phrase fragment	3.20%	17.14%	*the end of the*
	NP with other post-modifier fragment	2.31%	7.45%	*the fact that the*
	Other NPs (NP fragment)	4.91%	4.04%	*the most important thing*
	subtotal	10.42%	28.63%	
VP-based	Anticipatory it + VP/AP	10.00%	5.09%	*it is important to*
	Passive verb + PP	0.00%	0.68%	*can be seen as*
	Copula *be* + NP/AP	6.98%	3.98%	*is one of the*
	(VP+) *that*-clause fragment	2.49%	5.78%	*that it is not*
	(V/Adj+) *to*-clause fragment	6.45%	7.95%	*would be able to*
	Pronoun/NP (+auxiliary) + be (+ …)*	8.46%	13.23%	*they are going to*
	(auxiliary+) active verb	9.94%	7.27%	*have a lot of*
	Yes/no and *wh*-question fragment	1.30%	0.00%	*what do you think*
	Personal pronoun + lexical verb phrase	20.71%	5.16%	*I believe that the*
	(V+) *wh*-clause fragment	0.95%	0.87%	*don't know how*
	subtotal	67.28%	50.00%	
PP-based	PP with embedded of-phrase fragment	2.07%	7.45%	*in the case of*
	Other PP fragment	7.99%	8.39%	*at the same time*
	subtotal	10.06%	15.53%	
Others		12.24%	5.53%	*As you can see*

Note. * Underlined categories are the patterns more widely used in conversation (Biber et al., 1999)

As can be seen in Figure 1, there are substantive differences between the two corpora in the distributions of bundles across categories. While VP-based bundles account for almost 70% of NICKLE bundles, NP- and PP-based bundles take up only about 10% respectively. The figures are roughly 50%, 30%, and 15% in LOCNESS. In previous research, high percentages of NP- and PP-based bundles are considered to be the hallmarks of academic prose or professional writing and a high number of VPs is a more typical pattern found in spoken registers. For example, in Biber et al. (1999) the corresponding percentages from a large corpus of academic prose were 31% (VP-based), 30% (NP-based)[1], and 33% (PP-based). In this regard, the LOCNESS writers are shown to be closer to the professional academic writers in terms of structural patterns.

The subcategory where the two groups showed the greatest difference is the "personal pronoun + lexical verb phrase" frame. NICKLE contains a markedly high percentage of lexical bundles with personal pronouns (e.g., *I think that the*) compared to LOCNESS (20.71% vs 5.16%). The frequent use of personal pronouns is one of the common patterns found in Korean students' EFL argumentative writing (S. C. Hong, 2013; J. H. Kim, 2013), and this study once again confirms this pattern. Indeed, compared to LOCNESS, comprised of similar argumentative writing, NICKLE showed overall higher percentages in the subcategories that are more representative of speech registers (see underlined subcategories in Table 3). In addition, though not shown in Table 3, a considerable proportion of NICKLE bundles (about 19%) contain a contracted form, verb contraction (e.g., *it's possible that*) or *not*-contraction (e.g., *I don't agree*), which is another telling feature of informal register (Biber et al., 1999). However, bundles with a contraction take up a very small portion (3.4%) in LOCNESS.

Other subcategories revealing the Korean writers' salient differences from their NS counterparts, on the other hand, involve nominalization, which is often associated with academic prose (Biber et al., 1999). Table 3 shows that the Korean students used the "NP + post modifier" frame, especially "NP + *of*-phrase fragment" bundles (e.g., *the nature of the*) much less than the NS students (3.2% vs 17.14%). This trend can also be found in another *of*-phrase fragment frame ("PP + *of*-phrase fragment", e.g., *in the form of*), which occurred relatively rarely in NICKLE (2.07%). These two frames are identified in previous studies as extremely productive structures in academic prose (Biber, Conrad, & Cortes, 2003; Chen & Baker, 2010).

It should be noted, however, that the Korean students' writing showed a substantially greater use than the NS students (10% vs 5.1%) in one category exclusively associated with academic prose: "anticipatory *it*" (e.g., *it is important to*). "Anticipatory *it*" is usually

[1] This percentage of NP-based bundles are not directly comparable with that from this study as the former does not include "other NP fragments" (as presented in Chen & Baker, 2010)

exploited in academic prose to provide impersonalized evaluations (Ädel & Erman, 2012). Interestingly, while using a great number of bundles with personal pronouns, the Korean writers used at the same time an impersonal pattern frequently. It can be inferred that the Korean writers might have perceived (or explicitly learned) the pattern as a marker of formal writing and employed it as often as possible. Ädel and Erman (2012) indeed suggest that the structure can be a useful strategy for projecting a "detached writing persona" (p. 87), but the frequent use of the structure has also been problematized as a pattern non-native students use to emphasize propositions more than necessary (Hewings & Hewings, 2002).

All these findings presented above (i.e., the frequent use of the patterns characteristic of conversation, absence of the patterns typical of academic prose, and mixture of features of different registers) suggest that the Korean college EFL writers may have been at early stages of writing proficiency development, displaying lack of register awareness.

3. Discourse Functions

Results also revealed wide differences in the distributions of lexical bundles across discourse functions. Figure 2 below represents the distribution of bundles in each corpus across discourse functions and Table 4 breaks down the distributions across subcategories.

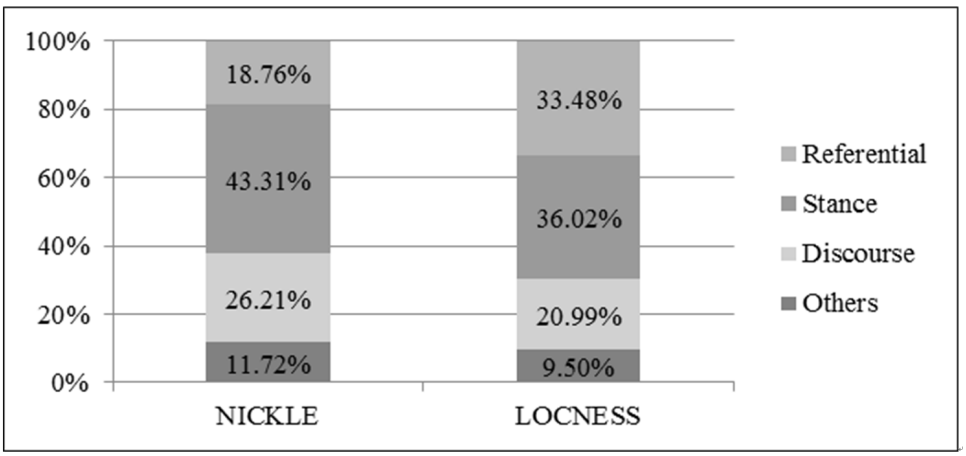

FIGURE 2 Distribution of Bundles Across Discourse Functions in Each Corpus (token)

TABLE 4

Distribution of Lexical Bundles across Subcategories of Discourse Functions (token)

Category	Subcategory	NICKLE	LOCNESS	Example
Referential expressions	Framing	3.91%	14.53%	*in the case of*
	Quantifying	10.59%	14.78%	*a large number of*
	Imprecision	1.01%	0.00%	*is a kind of*
	Place/time/text deitic	3.25%	4.16%	*for the first time*
	subtotal	18.76%	34.48%	
Stance bundles	Epistemic	15.86%	12.67%	*therefore I think that*
	Obligation/directive	10.12%	5.96%	*it is necessary to*
	Desire	4.67%	2.42%	*if you want to*
	Intention/prediction	0.30%	5.34%	*is not going to*
	Ability	12.37%	9.63%	*to be able to*
	subtotal	43.31%	36.02%	
Discourse organizers	Topic introduction	5.92%	3.17%	*let's think about*
	Topic elaboration/clarification	10.83%	5.16%	*on the other hand*
	Identification/focusing	7.57%	7.64%	*is one of the*
	Inferential/causative	1.89%	5.03%	*as a result of*
	subtotal	26.21%	20.99%	
Others		11.72%	9.50%	*that it is not*

Overall, in NICKLE, stance bundles take up the largest proportion (43.31%), followed by discourse organizers (26.61%) and referential expressions (18.76%). By contrast, LOCNESS showed a relatively even distribution of bundles across three functions with referential expressions being the largest category (34.48%). In previous research, conversation and academic prose have often been characterized by a predominance of stance bundles and referential expressions respectively (Biber et al., 2004; Chen & Baker, 2010). Thus, these results indicate that the Korean EFL writers' texts are closer to conversation rather than academic prose, which is consistent with the findings from the analysis of grammatical structures. This consistency can be explained by the fact that discourse functions seem to be performed to a large extent in certain grammatical structures, which is further discussed while looking into each category of discourse functions below.

First, among referential expressions, the two groups of writers showed the biggest difference in the use of framing expressions (3.91% vs 14.53%). Framing bundles are used to identify specific attributes or conditions and often take the grammatical structures of "NP + post modifier" (e.g., *the idea of a*) or "PP + of-phrase fragment" (e.g., *in the case of*), which are largely employed to abstract or conceptualize by nominalization what is being discussed, so typically used in academic writing. As noted above, the Korean writers rarely employed these frames.

Among stance bundles, epistemic bundles, which display knowledge about the following proposition or evaluate its (un)certainty, ranked first in both groups. This may not be surprising in that both corpora are of argumentative writing where writers express

their personal opinions. However, a close look at the actual epistemic stances employed reveal that the Korean writers' bundles have much more of speech-like lexical items such as "*I (don't) think*" and "*seem(s) to me*" (Paquot, 2010) than the native writers.

Turning to discourse organizers, the Korean writers employed a relatively high number of bundles that help to introduce and elaborate on a topic but did not use many inferential bundles. Of particular note is that while the NICKLE writers employed a high proportion of bundles for topic elaboration/clarification, the large majority of them are from a small set of logical connectives such as *on the other hand*, *at the same time*, and *that is to say*. These connectives were used frequently by the NS students as well, but the frequencies are markedly high in NICKLE. For example, the bundle *on the other hand* occurred 71 times in NICKLE, by far the most frequent of all bundles identified. Its usage in text will be discussed in greater detail later. The frequent use of adverbials including connectives within a short text is known to typify L2 writers' timed essays in which they have to demonstrate their ability to build logical relations within given time and length of text (Leedham & Cai, 2013; Paquot, 2010). Taken together, the results suggest that the Korean writers give the impression of logicality with the use of these bundles but without actually making logical inferences as evident in the low frequency of inferential bundles.

Finally, there is one issue often raised in previous studies with regard to discourse functions: cautious language or hedging manifested in bundles. Congruent with the findings of previous learner bundle studies (Ädel and Erman, 2012; Chen & Baker, 2010), less hedging was found in the bundles of NICKLE than of LOCNESS. While the Korean writers employed some of hedging language such as "*be likely to*" in their bundles, there were not many that can be seen as hedging bundles such as containing hedging verbs *(suggest, indicate, imply,* etc.), hedging nouns (*fact, assumption, possibility, evidence,* etc.) and degree nouns (*extent, degree* etc.). One striking absence is that of modal verbs. No bundles were found in NICKLE that contain modal verbs *would, could, might* or *can* (meaning possibility). LOCNESS, on the other hand, has a relatively high number of bundles with these modal verbs (e.g., *it would be a*, *could be used to*) with *would* appearing in 100 bundle tokens. Moreover, there are some bundles in NICKLE that can be seen as what Chen and Baker (2010) call over-generalizing or unnecessarily forceful bundles, such as *all over the world* (17 tokens), *people in the world* (8 tokens), and *the most important thing*[2] (21 tokens). With the salient lack of hedging language on the one hand, and the frequent use of overstating bundles (coupled with repeated use of "anticipatory *it*") on the other, the essays written by the Korean students may give the impression of being more

[2] In most instancees of *the most important thing,* the writer tends to present a single solution as the best for a given problem, somewhat too simplistically and categorically, leaving no room for alternatives.

categorical and forceful than is appropriate.

4. Overuse, Underuse, Mis-use of Lexical Bundles

TABLE 5 lists the most frequent 20 bundles from each corpus. Within the 20, both groups shared six bundles, which are underlined.

TABLE 5

The Most Frequent 20 Bundles from Each Corpus

NICKLE		LOCNESS	
Bundles	*f*	Bundles	*f*
on the other hand	71	*on the other hand*	30
don't have to	38	*one of the most*	28
I think it is	34	*is one of the*	26
we don't have	26	*in the case of*	26
don't want to*	25	*as a result of*	22
is one of the	24	*the end of the*	22
one of the most	23	*the fact that the*	22
for a long time	23	*the only way to*	19
at the same time	22	*have the right to*	18
I don't think	22	*to the fact that*	18
the most important thing	21	*do not* want to*	16
it is hard to	20	*a great deal of*	16
don't have any	20	*at the same time*	15
have the right to	19	*not be able to*	15
they don't have	19	*the rest of the*	15
but it is not	19	*when it comes to*	14
it is important to	17	*at the end of*	14
all over the world	17	*there would be a*	13
is a kind of	17	*would have to be*	13
so we have to	17	*will be able to*	12

Note. * In this study contracted forms are treated the same as their non-contracted forms in order not to inflate the number of bundle types (see Method for detail). Here, more dominant forms (i.e., contracted vs non-contracted) are presented for each corpus.

The list above nicely encapsulates what has been presented in the preceding sections for structural and functional analyses of the identified lexical bundles. The unique patterns identified above for each writer group are well manifested in the list. The Korean EFL writers relied much on bundles in the patterns of "personal pronoun + verb," and "anticipatory *it*," and showed a strong preference for contraction (6 out of 20 bundles). The NS students, meanwhile, made frequent use of framing bundles in the "NP + post modifier" frame, and bundles with an "*of*-phrase fragment" (8 out of 20). However, there are also six bundles shared by the two groups, most of which rank high in each corpus. Four out of the six shared bundles can be seen as discourse organizers (*on the other hand, is one of the, one of the most,* and *at the same time*). These bundles in fact seem to be

widely used in academic writing regardless of writing proficiency as they also make the lists of top bundles in other studies (Ädel & Erman, 2012; Chen & Baker, 2010; Staples et al., 2013; Wood, 2010). This suggests that though discourse organization is not the most frequent discourse function served by lexical bundles, some of the most frequent bundles are discourse organizers.

The list above is based on raw frequencies alone and so do not tell much about the distinctiveness of each bundle in NICKLE relative to LOCNESS as a reference corpus. As outlined in Method, log likelihood (LL) analysis helps to identify the bundles that are unusually frequent and unusually rare in comparison with the reference corpus, which can be regarded as overused and underused bundles respectively. Table 6 below lists the top ten bundles overused and underused by the Korean writers.

TABLE 6

Ten Bundles Overused and Underused by NICKLE Writers

Overuse				Underuse			
Bundles	N*f*[a]	L*f*[b]	LL[c]	Bundles	N*f*	L*f*	LL
the most important thing	21	0	32.87	*to the fact that*	0	18	21.99
don't have any	20	0	31.31	*the fact that the*	1	22	20.22
I think it is	34	5	29.46	*there would be a*	0	13	15.89
is a kind of	17	0	26.61	*would have to be*	0	13	15.89
so we have to	17	0	26.61	*both sides of the*	0	11	13.44
if we don't	16	0	25.05	*example of this is*	0	11	13.44
there are lots of	16	0	25.05	*the introduction of the*	0	11	13.44
on the other hand	71	30	24.92	*the end of the*	3	22	13.23
don't have to	38	10	22.58	*at the end of*	1	14	11.32
we don't have	26	4	22.03	*should not be allowed*	0	9	11.00

Note. a. frequencies in NICKLE, b. frequencies in LOCNESS, c. log likelihood ($p < 0.001$ for all bundles listed)

Confirming the findings discussed above, LL analysis reveals that overall the Korean writers overused an informal word (*thing*), bundles with first person pronouns (*I* and *we*) and contractions (e.g., *don't*), all of which can be seen as features of speech, while underusing NP and PP-based bundles, more typical of academic prose. One more finding of note is that the pattern of rhetorical overstatement typical of L2 writing combined with lack of hedging is confirmed in the Table 6. Along with *the most important thing*, the most overused bundle, *I think it is* can also be seen as an overstating bundle. Aijmer (2001) links the use of *I think* in L2 argumentative essays to non-native writers' attempt to make their argument more persuasive, which he sees is communicatively unnecessary. By contrast, many of the underused bundles can be associated with hedging (i.e., the hedging noun *fact* and modal verb *would*).

Among others, *on the other hand* is worth a detailed discussion as it ranks first in frequency in both corpora (see Table 5) but still finds itself among overused bundles. In

fact, its frequency is much higher in NICKLE, over twice that in LOCNESS. One possible reason why the Korean writers relied so heavily on this specific bundle is that they may have used it as one of the "lexical teddy bears" they felt familiar and safe to use (Hasselgren, 1994), which is not surprising as non-native writers at early stages of their writing development may lack phrasal repertoires at their disposal. Then, more important may be whether the Korean writers used it appropriately in context. A qualitative perusal of concordance of this bundle from both corpora may reveal the differences in its use, which is discussed below.

On the other hand (OTOH) is in fact one of the most frequently used and widely researched discourse markers. While there are several categorizations of OTOH uses in discourse, Bell (2004) offers a relatively simple classification. He classifies its uses broadly into three categories: listing, contrastive, and cancellative. OTOH as a listing marker lists two items mostly used together with its correlative adverbial *on (the) one hand*. OTOH as a contrastive marker compares and contrasts two items often implying alternativity. Lastly, OTOH as a cancellative marker cancels or mitigates an aspect of the preceding proposition, often replaceable with markers such as *but, however,* and *yet*. Based on a large mixed-genre corpus of native speaker discourse, Bell identified contrastive use as the most common in academic writing. He also found that OTOH in the sentence-initial position is likely to be cancellative while post-NP OTOH tends to be contrastive, serving as a focus marker. Table 7 shows the distributions of OTOH uses in NICKLE and LOCNESS.

TABLE 7

Uses of *on the other hand* in Two Corpora

Categories	NICKLE		LOCNESS	
	f	%	*f*	%
Total No. of tokens	71		30	
Contrastive	59	83.1%	23	76.7%
Cancellative	7	9.9%	5	16.7%
Listing	5	7.0%	2	6.7%
Sentence initial	62	87.3%	18	60.0%
Sentence medial (post-NP)	9 (3)	12.7% (4.2%)	12 (9)	40.0% (30.0%)
on (the) one hand	0	0.0%	2	6.7%

As can be seen in Table 7, a great majority of OTOH tokens in both corpora were of contrastive use, but the LOCNESS writers were slightly more varied in its distribution across the three categories and also had two instances of *on (the) one hand* used together with OTOH. In terms of the position of OTOH, however, substantial differences emerge between the two groups. As analysis of all instances of OTOH goes beyond the scope of this paper, only the sentence-medial OTOHs are discussed below.

In NICKLE, the bundle is used in the sentence-medial position only 12.7% of the time

while 40% of OTOH tokens in LOCNESS occur in the sentence-medial position. As noted above, Bell (2004) generalizes from his corpus analysis that contrastive use of OTOH is usually realized when OTOH occurs in the sentence-medial, especially post-NP position, operating as a focus marker. Although a majority of contrastive uses in LOCNESS still occur in the sentence-initial position, nine of 12 sentence-medial OTOH tokens (30% of the total) occur in a post-NP position as in the example below and they all are of contrastive use.

(1) Unorganized crime does not pay, at least not very well. Organized crime, *on the other hand*, pays off quite handsomely. (LOCNESS)

OTOH in (1) functions as a focus marker to emphasize the preceding NP (*organized crime*) as the item being compared (or contrasted) to the prior discourse. In NICKLE, by contrast, only nine OTOH tokens (about 13%) occur in the sentence-medial positions and only three tokens of them (4.2% of the total) are post-NP, functioning as contrastive.

(2) In the case of the flu, one usually gets better in about a week. This disease, *on the other hand,* gets worse. (NICKLE)

Most of the other sentence-medial OTOHs in NICKLE are used erroneously or ineffectively as a contrastive marker as in (3) and (4) below.

(3) In industrialized society, most people are pressed by time so that they have no time to have a meal with their composure. It needs much time to prepare Korean traditional food made from rice *on the other hand*, preparing western food made from flour is convenient and needs little time. (NICKL)
(4) Moreover most Korean schools don't have any lawn field in which young soccer players can practice and some schools even hold no playgrounds, *on the other hand* more than 80% Japanese schools possess their own lawn fields. (NICKLE)

OTOH in (3) is erroneously used as it occurs at the end of the first proposition (i.e., Korean food taking long to cook), rather than with the usual second proposition (i.e., western food being quick to prepare), rendering the meaning of the whole paragraph rather confusing. By contrast, (4) is clear in its meaning and OTOH here is obviously of contrastive use. Perhaps due to the writer's lack of syntactic control, however, OTOH is used like *while*, a subordinate conjunction, resulting in a run-on sentence. Breaking the sentence into two and placing OTOH after the NP of the second sentence (i.e., Japanese schools) would produce syntactically correct and rhetorically more effective discourse.

As seen above, a qualitative analysis of OTOH reveals the unique uses, many of the being erroneous and inappropriate, by the Korean EFL college writers, showing quite deviations from the native speaker norm. This small-scale analysis of OTOH serves as one good example showing that the overuse of certain lexical bundles can go hand in hand with misuse.

V. CONCLUSION

The present study investigated the characteristics of four-word lexical bundles from a corpus of EFL argumentative essays written by Korean university students in comparison with those of English L1 students. In so doing, the study employed rigorous procedures in its identification and analysis of lexical bundles. Confirming the findings of previous learner bundle research on Korean EFL learners' argumentative writing, the lexical bundles retrieved from NICKLE were in large part the ones widely used in speech registers. Specifically, NICKLE bundles are characterized by high instances of personal pronouns, contractions and stance bundles on the one hand, and lack of nominalizations (NP, PP phrases), hedging and referential bundles on the other. These patterns found in NICKLE bundles showed substantial differences from those of LOCNESS, which were more typical of academic prose. One exception found, however, is the highly frequent use of "anticipatory *it*", which is usually associated with formal writing. Combined with other highly frequent uses of linking adverbials (e.g., *on the other hand*) and overstatement (e.g., *I think*, and *all over the world*), the results suggest that while generally lacking register awareness, the Korean writers may have relied heavily on what they believed to be markers of argumentative writing. Meanwhile, a qualitative analysis of the most frequent bundle in both corpora, *on the other hand*, demonstrated that the Korean writers deviate from the native speaker writers not only in the frequency and distribution of lexical bundles but also in the ways they are used in text.

These findings carry important implications for the pedagogy of EFL composition in Korea. Along with the previous studies, the present study demonstrated Korean university students' overall lack of register awareness in their English argumentative writing. This can be understandable in that the type of argumentative composition Korean university students usually engage in (either in EFL composition classes and for standardized tests like TOEFL and TOEIC) is short (4-5 paragraphs) timed essays where they have to provide personal opinions on some popular topics. The writers should display their ability to be persuasive and to control logical relations across different sections in a short text within a tight time limit. Without proper exposure to or explicit learning of genre features of argumentative writing and broader academic writing, they may resort to the use of personal

pronouns (believing that makes their argument more persuasive) and highly repeated use of a limited set of phrases and patterns they feel safe to use in academic writing. Specifically, they can be guided, for example, to make less first-person references such as *I think* and *in my opinion* but to employ more hedging devices marking degrees of probability such as *would, may*, and *be likely to* instead.

This means that Korean university students and other adult learners should be given more opportunities to familiarize themselves with argumentative writing that go beyond short timed essays and acquire competence in using genre features and conventions of academic writing in their EFL courses (H. S. Park, 2012) or better English for Academic Purposes (EAP) courses. Phraseologies including lexical bundles should be part of the competence. One promising way of improving phraseological competence would be to take advantage of the growing availability of corpora of NS student writing[3] like the one used in this study. The native students' writing has greater similarities in topics and sophistication in content organization to Korean students' writing than published journal articles, which have often been used as models for student writing. Using these corpora is likely to provide the types and usages of formulaic language that are more accessible and relevant to Korean EFL student writers.

Although shedding new light on lexical bundles from Korean university students' argumentative writing, the present study has limitations as well. One major limitation would be that the study looked at only four-word strings, which only gives a partial picture of formulaic language and other textual features occurring in NICKLE and so the conclusion on its register characteristics is only partial too. Another limitation is the classification of bundles. Ädel and Erman (2012) pointed out cogently that the analytic framework based on Biber et al. (1999) leaves much room for arbitrary interpretation and application. We, the two authors, also had considerable difficulty in consistently applying the framework and arrived at decisions in not a few cases that ran counter to the classifications of previous studies. Lastly, although we emphasized the need for complementing a frequency-based study with qualitative investigations, the textual analysis done for the study was far from comprehensive.

We may thus conclude by suggesting some future studies drawn from the acknowledged limitations. Going beyond quantitative tallying of bundles across functional and structural categories, future studies can complement quantitative analysis with in-depth qualitative explorations of the ways specific items are used in text. The mixed analysis can further be complemented with interviews with the essay writers to investigate the intentions and

[3] The British Academic Written English (BAWE) corpus (http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/) and the Michigan Corpus of Upper-level Student Papers (MICUSP) (http://micusp.elicorpora.info/) are two major publicly available corpora of NS student writing.

motivations behind the use, overuse, and underuse of certain phraseologies. These studies combined will greatly inform the pedagogy of EFL composition, and EAP instruction.

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*(2), 81-92.

Aijmer, K. (2001). I think as a marker of discourse style in argumentative Swedish student writing. In Aijmer, K. (Ed.), *A wealth of English: Studies in honour of Göran Kjellmer* (pp. 247-257). Göteborg, Sweden: Acta Universitatis Gothoburgensis.

Altenberg, B. (1998). On the phraseology of spoken English: the evidence of recurrent word-combinations. In A. P. Cowie (Ed.), *Phraseology: theory, analysis and applications* (pp. 101-122). Oxford: Oxford University Press.

Bell, D. M. (2004). Correlative and non-correlative "on the other hand". *Journal of Pragmatics, 36*(12), 2179-2184.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics, 14*(3), 275-311.

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

Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: an initial taxonomy. In A. Wilson, P. Rayson & T. McEnery (Eds.), *Corpus linguistics by the Lune: a festschrift for Geoffrey Leech* (pp. 71-93). Frankfurt, Germany: Peter Lang.

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

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

Chen, Y.-H., & 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.

Dunning, Ted. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, *19*(1), 61-74.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 145-160). Oxford: Oxford University Press.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237-258.

Hewings, M., & Hewings, A. (2002). 'It is interesting to note that': A comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes, 21*(4), 367-383.

Hong, Shin-Chul (2013). An n-gram analysis of Korean English learners' writing. *Korean Journal of English Language and Linguistics 13*(2), 313-336.

Hong, Shin-Chul (2015). The comparison of n-gram use between intermediate and advanced Korean learners of English. *Journal of Language Sciences 22*(1), 147-170.

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

Kim, Junghyun (2013). Lexical bundles in Korean college students' English essays: A corpus-based comparative study. *English Language & Literature Teaching, 19*(3), 157-179.

Kwon, Ye-Eun & Lee, Eun-Joo. (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.

Lee, Eun-Joo (2009). A corpus-based study of the Korean EFL learners' use of formulaic sequences. *Foreign Languages Education, 16*(2), 321-340.

Leedham, M., & Cai, G. (2013). Besides… on the other hand: Using a corpus approach to explore the influence of teaching materials on Chinese students' use of linking adverbials. *Journal of Second Language Writing, 22*(4), 374-389.

Paquot, M. (2010). *Academic vocabulary in learner writing: From extraction to analysis.* London, UK: Continuum.

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.

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

Park, Hyesook (2012). Genre-based instruction and Korean college students' development of expository writing in English. *Modern English Education, 13*(1), 43-67.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* Houndmills, UK: Palgrave Macmillan.

Wood, D. (2010). Lexical clusters in an EAP textbook corpus. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 88-106). London, UK: Continuum.

Examples in: English
Applicable Languages: English
Applicable Levels: Elementary

Choongil Yoon
Department of English Education
Ewha Womans University
52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760 Korea
Tel: H.P.: 010-6476-0807
Email: chongal2@hotmail.com

Ji-Myoung Choi
Dept. of Language and Information
Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul 120-749
Tel: H.P. 010-4352-4594
Email: amancio.choi@gmail.com

Received 15 June 2015
Revised 4 August 2015
Accepted 17 August 2015