Linguistic Research 35(Special Edition), 105-135
DOI: 10.17250/khisli.35..201809.005

Lexical bundles in ESP writing: Marine accident investigation reports*

Se-Eun Jhang · Sungkuk Kim · Yilian Qi**
(Korea Maritime and Ocean University·Sungkyunkwan University·Dalian Maritime University)

Jhang, Se-Eun, Sungkuk Kim, and Yilian Qi. 2018. Lexical bundles in ESP writing: Marine accident investigation reports. *Linguistic Research* 35(Special Edition), 105-135. Previous research on lexical bundles produced by native versus non-native English speakers has been mostly carried out within the academic domain, yet it is not fully understood in ESP context. This study investigates the construct of lexical bundles in the genre of marine accident investigation reports (MAIR).Through comparison of lexical bundles used by L1-English versus L1-Japanese professionals in MAIR, differences between two groups are clearly displayed. It is found that compared with English reporters, Japanese professionals employ a considerably wider range of four-word bundles, exhibit an overuse tendency in almost all structural patterns and functional types and adopt different strategies to construct lexical bundles and fulfill discourse functions. Some similarities are also discovered between the two groups of writers, which are believed to reflect the special characteristics of MAIR genre. **(Korea Maritime and Ocean University·Sungkyunkwan University·Dalian Maritime University)**

Keywords lexical bundles, ESP writing, marine accident investigation reports, native and non-native English speakers

1. Introduction

As a type of multi-word units retrieved through a frequency-driven approach, lexical bundles are recurrent continuous word sequences that largely straddle the boundary between lexis and syntax, functioning as "basic building blocks of discourse" (Biber and Barbieri 2007: 270; Biber, Conrad, and Cortes

* The earlier version of this article was presented at the summer conference held in Pusan National University, August 17, 2017 by the Korean Association of Language Sciences. The authors would like to thank the audience as well as two anonymous reviewers and Professor Robert Dickey for their valuable comments. All errors in this article are ours.

** Se-Eun Jhang is the first author, Sungkuk Kim is the co-author, and Yilian Qi is the corresponding author.

2004: 371). The prevalence of lexical bundles in discourse has prompted a multitude of studies with various research interests, including the studies of lexical bundles and disciplinary variation, such as disciplines from pure sciences and social sciences (Hyland 2008b); comparison of lexical bundles in different discourse contexts, such as textbooks versus classroom discourse (Biber and Barbieri 2007; Biber, Conrad, and Cortes 2004); and the production of lexical bundles among different populations, such as learners versus experts (Cortes 2004; Hyland 2008a) and native speakers versus non-native speakers (Chen and Baker 2010; De Cock 2000; Römer 2009). For the analysis of lexical bundles produced by native versus non-native English speakers, a review of relevant literature shows that most studies have been carried out within academic prose. These studies have found that there exist great differences between two groups of writers in their academic written work, although the extent of the differences varies due to different research designs (Chen and Baker 2010; De Cock 2000; Erman 2009; Granger 1998; Howarth 1998; Lewis 2009; Römer 2009). However, very little research on lexical bundles analysis in native/non-native writing exists in the ESP context. We suggest that it is valuable to conduct such a study, given the fact that English, as a leading lingua franca, has been used as a device for a wide range of professions and it is a common practice that many written documents are produced by non-native speakers of English. Understanding lexical bundles constructed by both native and non-native English speakers in ESP genre-based writings not only provides evidence on variations in language use, but also helps better understand the ESP genre in which these writers participate. Therein lies the intended contribution of the current research. In this study, we chose the marine accident investigation reports (MAIR) from the maritime domain as the target ESP writing, and put particular focus on comparing the use of lexical bundles by L1-English versus L1-Japanese professionals. Through qualitative and quantitative comparison of the overall structural and functional patterns of lexical bundles employed by these two groups, this study set out to explore whether the use of lexical bundles by Japanese writers deviates from native English speakers' norms in MAIR. Thus, we investigate the following research questions:

1) What are the differences between lexical bundles used by L1-English versus L1-Japanese professionals in MAIR in terms of structural types?
2) What are the differences between lexical bundles used by L1-English versus L1-Japanese professionals in MAIR in terms of discourse functions?

By answering the above questions, we hope to raise the readers' awareness on the existence of varieties within the MAIR genre, and more importantly, gain insights into the formulaic nature of MAIR discourse, which can serve as a starting point for learning and teaching practice.

2. Previous research on the use of lexical bundles by native vs. non-native English speakers

2.1 Studies conducted in the EAP context

The researchers' interest in the use of lexical bundles by native and non-native English speakers has been greatly inspired by one of Sinclair's earliest studies (1991), where he found that native speakers are highly dependent on the use of prefabricated word chunks in writings while non-native speakers showed lack of phraseological capabilities. Such a finding was further confirmed by multiple subsequent studies which have been launched either within various discourse contexts or with human subjects from different language backgrounds. For instance, De Cock (2000) analyzed the construction of lexical bundles by non-native English speakers from a perspective of second-language learning and found that L2 users of English often rely on L1 transfer in lexical bundle constructions, which results in misuse of lexical bundles in the case that there is no match between L1 and L2, and overuse of constructions with shared L1 equivalents. Unlike De Cock, Chen and Baker (2010) compared the use of lexical bundles by native and non-native speakers without considering the issue of language transfer. Instead, they carried out a 3-way comparison among L1-English students, L1-English experts and L2-English students. Their results indicated that L2 students exhibited a tendency towards overusing and underusing certain types of lexical bundles that are typical in academic prose,

although there were relatively fewer differences between L1 and L2 student writing than the differences between students and experts. Other researchers have also undertaken studies of lexical bundles focusing on L1 versus L2 distinction, most of which have found that the overuse, underuse and misuse of lexical bundles are characteristics of L2-English writing (e.g. Ädel and Erman 2012; Cortes 2004; Nekrasova 2009; Pan, Reppen, and Biber 2016; Schmitt 2004). From the above, we can see that the significance of competence in using lexical bundles for both native and non-native speakers has been highly emphasized in academic genre.

2.2 Studies conducted in the ESP context

Similar to academic writings, it is also important to understand the nature of how lexical bundles are used by authors from different linguistic backgrounds in ESP genre-based writings. However, the existing literature shows that it has not received much attention and importance, as compared to the same type of research in EAP genre. To date, investigation of lexical bundles in the ESP context can be found in research such as Jablonkai (2010), Breeze (2013), Jhang and Lee (2013) and Grabowski (2015), among others. These studies were primarily designed to identify the characteristics of lexical bundles in a particular genre or text type rather than look into the variations in use of lexical bundles produced by different groups of writers. To be specific, Jablonkai (2010) explored lexical bundles in EU documents. Jhang and Lee (2013) and Lu, Lee, and Jhang (2017) analyzed clusters and key clusters in a Maritime English corpus. The study of Grabowski (2015) provided insights into the constructs of lexical bundles within English pharmaceutical discourse, while Breeze (2013) merely attempted to understand the nature of lexical bundles in four different legal genres without discussing the impacts that the writers' linguistic backgrounds had on the use of lexical bundles.

Drawing on the previous research, the present study seeks to fill this research gap through comparison of lexical bundles constructed by English versus Japanese professionals in MAIR. The reasons for choosing MAIR as the target ESP genre-based writing will be given in the following section.

3. Data and methodology

MAIR is an essential written text type among all the maritime-related writings, since safety issues are always one of the greatest concerns in maritime domain. To prevent and avoid marine accidents, MAIR is required to be provided in each accident investigation. Overall, it functions as a platform where experts can report investigation findings, explain the causes of the accident and express recommendations for other vessels. Close observation of MAIR in different cultural contexts points out its various aspects and dynamic nature. In another words, the MAIR conducted by professionals from different countries displays notable differences in terms of report format, length, narrative styles and linguistic features, etc. Therefore, the value of comparing the use of lexical bundles across the subsets of the MAIR genre lies not only in providing an overall understanding of MAIR discourse, but it also raises the report readers' awareness of the varieties existing within this genre.

3.1 Two study corpora

Two MAIR corpora were compiled for the purpose of this study. One corpus, labeled MAIR-EN, consists of British marine accident investigation reports, representing native English writings, as it has been commonly recognized as the standardized format for MAIR. These documents can be freely accessed from the official websites of the U.K. Government (https://www.gov.uk/maib-reports). Another corpus, labeled MAIR-JP, is comprised of English marine accident investigation reports written by native-Japanese professionals. The data in the MAIR-JP corpus was derived from the entire collection of the English reports available on the official website of Japan Transport Safety Board (JTSB, https://www.mlit.go.jp/jtsb/marrep.html), all of which were explicitly identified as the English translations of the Japanese original investigation reports.[1] Therefore, these could be considered representative of the Japanese writers' productions.[2] Additionally, since the data in the MAIR-JP corpus were chosen

1 As noted in the JTSB website, the English version report has been translated and issued by JTSB to make its reading easier for those English speaking people who are not familiar with Japanese.

2 As Professor Robert Dickey at Keimyung University, proofreading the earlier versions, pointed out

from the reports which have been published during 2009-2016, the selection of the counterparts for the MAIR-EN corpus is thus confined to the same time period to ensure its comparability with MAIR-JP corpus. For accurate data processing, all the selected reports of both corpora were converted into plain text files and cleaned of headings, formatting, diagrams, images and appendices.

Based on the above criteria, the final MAIR-JP corpus contains 56 reports with 733,708 running words and the size of MAIR-EN corpus is around 1.85 million words, covering 194 reports. The detailed contents of each corpus and their respective word counts are outlined in Table 1.

Table 1. Constituents of two MAIR corpora

Corpus	Representation	Data Source	Number of reports	Total number of words
MAIR-EN	English reports written by British speakers	U.K.	194	1,852,552
MAIR-JP	English reports written by Japanese speakers	Japan	56	733,708

3.2 Size of corpora

As shown in Table 1, both study corpora are not large in size, this is especially true with the MAIR-JP corpus. Despite their small scales, we can assume that the sizes are sufficiently suitable and the two study corpora are representative for investigating the use of lexical bundles in the MAIR genre respectively because descriptive linguistics should not be intimidated by the 'need' for larger corpora (Biber 1990). Rather, smaller corpora are more suitable than large multi-million word corpora to identify linguistic patterns in ESP contexts (Grabowski 2015; Koester 2006).

It is also noticeable that there exists a disparity in size of two study corpora. The MAIR-EN corpus is much larger than MAIR-JP corpus. Therefore, the frequency of lexical bundles in both corpora was normalized to a rate per one

---

that the English reports in MAIR-JP are only the translation versions of Japanese reports written by translators but not maritime professionals, this claim could be problematic. But we will not go into this discussion here.

million words for comparison across two corpora.

3.3 Analysis procedures

3.3.1 Lexical bundle identification

The first step of the analysis was to generate a list of lexical bundles in both of the study corpora. This selection process is guided by several key criteria, namely, the length of bundles, the cut-off frequency criterion and the dispersion threshold. As for the length of lexical bundles, only the four-word bundles were considered in the present study. This is partly because the four-word scope offers a more readily recognizable range of structures and functions, which could be good discriminators of registers (Cortes 2004; Hyland 2008; Scott and Tribble 2006). Another reason is that it is the most favored length for writing studies (Chen and Baker 2010). Its prevalence therefore allows us to compare our data with that used in other genres, such as academic prose (Biber and Barbieri 2007; Biber, Conrad, and Cortes 2004; Cortes 2004; Hyland 2008). All the candidate four-word combinations were automatically retrieved by the cluster setting function of Wordsmith 6.0 software (Scott 2016).

With regard to the frequency threshold, the cut-off point for our study is set at 40 times per million words, a moderately high frequency threshold used in most of the previous lexical bundle studies (Biber and Barbieri 2007; Bernardini, Ferraresi, and Gaspari 2010; Gaspari 2013; Goźdź-Roszkowski 2011; Jukneviciene 2009; Pan, Reppen, and Biber 2016). This standardized frequency is equivalent to a raw frequency of 74.1 times in the MAIR-EN corpus and 29.3 times in the MAIR-JP corpus, as shown in Table 2. As two corpora in this study are different in size, there exists possibility that "the cut-off frequency would lose its expected impartiality after being converted into raw frequencies" (Chen and Baker 2010: 32). In order to prevent loss of impartiality, we then rounded down these two numbers to 74 and 29 respectively and converted them into normalized frequency again. It was confirmed that the corresponding normalized frequencies after rounding were the same as the originally reported frequency threshold (39.6 and 39.9 can both be rounded up into 40), as shown in Table 3; therefore, it

could be argued that the 40 times per million words, which we set as the standardized cut-off frequency, operates well in the present study.

Table 2. Normalized and corresponding raw frequency thresholds for comparison

Corpus	Set normalized frequency threshold (per million words)	Corresponding raw frequency
MAIR-EN	40	74.1
MAIR-JP	40	29.3

Table 3. Raw and corresponding normalized frequency thresholds adopted

Corpus	Set Raw Frequency threshold	Corresponding normalized frequency(per million words)
MAIR-EN	74	39.9
MAIR-JP	29	39.6

As the last criterion for lexical bundle identification, dispersion threshold is used to avoid idiosyncrasies from individual writers/institutions (Biber, Conrad and Cortes 2004). In this step, we followed Hyland's (2008a) observation that the lexical bundles have to occur in at least 10% of all texts in the corpus. Therefore, we identified all the four-word bundles occurring in more than 19 different texts from the MAIR-EN Corpus and 5 texts from the MAIR-JP corpus.

3.3.2 Filtering out process

The second step of the analysis dealt with domain-specific and overlapping bundles, since the presence of these items has been considered to "inflate the results of quantitative analysis" (Chen and Baker 2010: 33). According to Chen and Baker (2010), overlapping bundles could be categorized into two types: (a) complete overlap and (b) complete subsumption. The term "complete overlap" denotes that two overlapping 4-word sequences that shared the same occurrences are indeed derived from one extended 5-word combination; another situation is that two or more overlapping bundles occur with varying frequencies, but the

occurrence of one of the bundles subsumes others. Thus bundles occurring in such an occasion are defined as "complete subsumption". For each type of the overlapping bundles, it is suggested to combine them into one longer unit in order to guard against inflated results. Following their suggestion, we identified overlapping bundles in each corpus. For example, in the MAIR-EN corpus, issues directly contributing to and safety issues directly contributing both occur 57 times per million words and are derived from a 5-word combination safety issues directly contributing to. In this case, these two overlapping bundles were replaced by the longer unit, which occurrences were counted as 57 per million words. After manually checking the concordance lines of each bundle in question, the merging process in the current study results in 15 exclusions from the bundle list of the MAIR-EN corpus and 13 from the MAIR-JP list.

As for domain-specific bundles, the decision made here is different from other research in which elimination has been recommended (Chen and Baker 2010; Hyland 2008; Pan, Reppen and Biber 2016). In this study, we decided to keep these bundles in our lists rather than omit them. This is because these bundles convey a range of grammatical structures and discourse functions that can reflect the specificity of the genre and also give valuable clues to the differences between two corpora (e.g. *the vessel was hit* in the MAIR-JP; *reproduced from admiralty chart* in the MAIR-EN, etc). Hence the minimal revision of the bundles can help keep it as authentic as possible. The results of manual filtering, including the numbers of lexical bundles before and after refinement, are listed in Table 4.

Table 4. Number of bundles (types and tokens) before and after refinement

Corpus	Before refinement		After refinement	
	No. of lexical bundles(types)	No. of lexical bundles(tokens)	No. of lexical bundles (types)	No. of lexical bundles (tokens)
MAIR-EN	149	20306	134	18421
MAIR-JP	443	34448	430	33405

3.3.3 Inter-rater reliability

3.3.3.1 Inter-rater reliability for filtering out process

Since filtering is operated by personal judgment, different views would inevitably occur during this process. For inter-rater analysis, the kappa statistic was chosen to measure agreement between the two researchers. It is shown that there are high degrees of agreement between two raters on reserving or removing certain bundles in both corpora (the kappa value is 0.91 for the MAIR-EN corpus and 0.89 for the MAIR-JP corpus), which implies that the two researchers were highly consistent with the initial bundle lists generated during this process. In cases of disagreement, researchers negotiated each case until they reached full agreement.

3.3.3.2 Inter-rater reliability for qualitative analysis

Once bundle lists were finalized, the last step was qualitative investigation of lexical bundles, including both structural and functional analysis. Again, the structural and functional types of lexical bundles were manually classified by the two researchers. The ratings of all classifications were aggregated and subjected to statistical analyses in order to assess the inter-rater reliability. The kappa values in both situations are > 0.75 (0.85 for structural classification and 0.81 for functional classification), which fall within a satisfactory level of reliability. Similarly, researchers discussed each case of disagreement to reach full agreement.

3.3.4 Statistical tests for lexical bundle comparison

One statistical test employed in this study is the log-likelihood (LL) statistic (Paul Rayson's online log-likelihood calculator is available on the UCREL[3] website at http://ucrel.lancs.ac.uk/llwizard.html).

Specifically, it is performed to determine the statistical significance of

3 UCREL stands for University Center for Computer Corpus Research on Language.

differences in each structural and functional category across two corpora. Although statistics such as LL test and chi-square test are both ''useful for comparing the relative frequency of words or phrases'' across corpora (Simpson-Vlach and Ellis 2010: 492), LL test has some preferable advantages. First, it does not necessarily require that the corpus data are normally distributed, which is the case of natural language (McEnery, Xiao, and Tono 2006). Second, it has been empirically proved as an effective measure for finding terms with low frequency in a corpus, yet in these cases chi-square test is invalid (Daille 1995; Dunning 1993). Based on the above, we believe that it is appropriate to adopt the LL test in our study since the counts of lexical bundles for certain types are low, such as lexical bundles functioning as text deictic in the MAIR-JP corpus and subject-specific bundles referring to equipments in the MAIR-EN corpus.

In addition, the standardized residual method is also adopted in this study with the purpose of identifying which functional types make a statistically significant contribution to the differences across the two corpora. The value of standardized residuals (R) is calculated in a chi-square contingency table, where the residual (difference between the observed and expected count of each cell) is divided by its standard deviation. By doing the calculations above, this measure is believed to discover "which cells contribute the most, and which contribute the least" (Lamart 2013). In the present study, this step was undertaken through SPSS version 21.0 (IBM Corp 2012) and the results are presented in the next section.

4. Results and discussion

4.1 Lexical bundle Lists in MAIR-EN versus MAIR-JP corpus

The above criteria yielded 134 four-word bundles in the MAIR-EN corpus and 430 counterparts in the MAIR-JP corpus, whose smaller dataset generates more lexical bundles. The result seems to indicate that the Japanese professionals use a considerably wider range of four-word bundles than do the English professionals. However it needs to be interpreted with caution, since one

possible reason lies in the discrepancy of corpus size. As reported by Chen and Baker (2010), a large corpus usually elicits higher converted raw frequencies and wider distribution which could lead to less retrieval of recurrent multi-word sequences. Therefore, the number of lexical bundles extracted from the MAIR-EN corpus is much less than that from the MAIR-JP corpus because of large corpus size.

A comparison of two bundle lists also allows us to discern that there are 36 lexical bundles shared by both groups of writers. This implies that the way English authors use lexical bundles is, to a large extent, different from Japanese counterparts. Except for this, no other comparisons of the bundle lists were undertaken at this step, as it is suggested that comparisons would be better made on the level of bundles' structural and functional characteristics instead of the direct comparisons of any specific bundle lists (Pan, Reppen, and Biber 2016). A reason for this is that the structural and functional features of lexical bundles are less influenced by the corpus designs and identification procedures, and hence can provide more valuable insights. Enlightened by this previous research, the retrieved bundles were subjected to structural and functional analyses and the following sections mainly demonstrate the differences in these two aspects.

4.2 Comparison of structural types of lexical bundles across two corpora

4.2.1 Comparison of distribution of structural categories across two corpora

The structural analysis was based mainly on the taxonomy proposed by Biber Joansson, Leech, Conrad, and Finegan (1999) for describing structural correlates of lexical bundles in academic prose. It was then supplemented by the "verb phrase with active verb" category taken from conversation register (Biber et al. 1999) to more fully represent the patterns that emerged from the data in Table 5 below. In order to provide a clear route for discussion, this categorization scheme (14 categories in total) was further grouped into three broader categories: "NP-based", "VP-based" and "PP-based" following Chen and Baker (2010). NP-based and PP-based bundles include noun phrases and prepositional phrases,

while VP-based bundles refer to "word combinations with a verb component" (Chen and Baker 2010: 35).

Table 5 presents the distributions of bundle tokens across structural categories.

Table 5. Distribution of structural categories across two MAIR corpora

Category	Structural pattern	Lexical bundle tokens (raw/normalized frequency per million words)	
		MAIR-EN	MAIR-JP
NP-based	1) noun phrase with *of*-phrase fragment *(time of the accident)*	4806 (2594)	6409 (8735)*
	2) noun phrase with other post-modifier fragment *(accidents occurring in the)*	1415 (764)	2302 (3137)*
	3) other noun phrase expressions *(contributory causes and circumstances)*	2780 (1501)	2033 (2771)*
PP-based	4) prepositional phrase with embedded *of*-phrase fragment *(as a result of)*	2935 (1584)	3058 (4168)*
	5) other prepositional phrase fragment *(as a basis for)*	2687 (1450)	5493 (7487)*
VP-based	6) anticipatory *it* + verb phrase/adjective phrases *(it is possible that)*	593 (320)	3208 (4372)*
	7) passive verb + prepositional phrase fragment *(identified during the investigation)*	547 (295)	2259 (3079)*
	8) copula *be* + noun phrase/ adjective phrase	-	-
	9) (verb phrase+) *that* clause fragment *(that the vessel was)*	319(172)	5734 (7815)*
	10) (verb/adjective+) *to*-clause fragment *(to comply with the)*	741 (400)	**257 (350)**
	11) adverbial clause fragment *(when the vessel was)*	196(106)	431 (587)*
	12) pronoun/noun phrase + be (+…) *(there was no evidence)*	350 (189)	837 (1141)*
	13) VP with active verb *(arrived on the bridge)*	889 (480)	1217 (1659)*
	14) Others *(the chief officer had)*	163 (88)	167 (228)*
Total		18421 (9943)	33405 (45529)*

As shown in Table 5, there are great differences between the two corpora in

the use of structural patterns of lexical bundles. The log-likelihood test comparing tokens indicated that 13 out of 14 structures were statistically overused by Japanese reporters relative to their English counterparts. But in the case of 'to-clause fragment' (category 8), no significant difference was found between these two groups.

When looking at the proportional distributions across three structural categories, differences between two groups of authors could also be detected. Figure 1 plots the percentages of each structural category in each corpus.

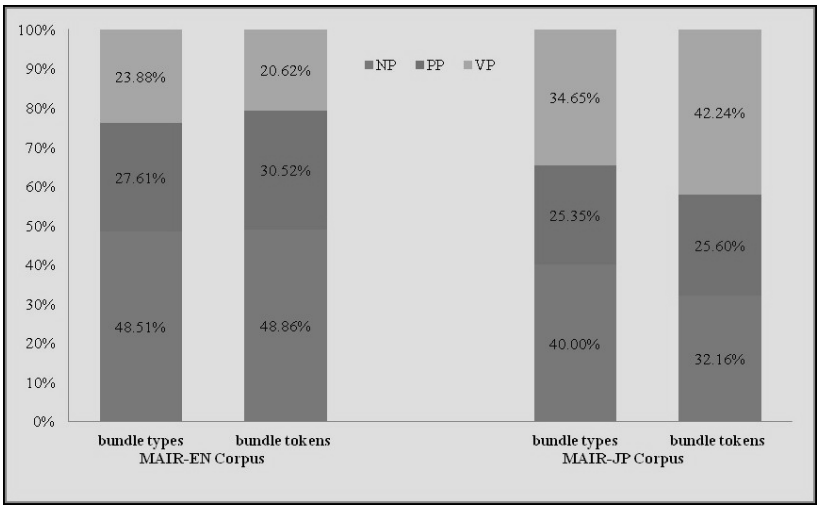

Figure 1. Proportional distribution of lexical bundles across three main structural categories in two MAIR corpora (type & token)

As we can see, there is dense use of NP-based bundles in the MAIR-EN corpus (i.e. 48.51% of bundle types and 48.86% of bundle tokens), suggesting that this structure is predominantly used by English professionals. However, the MAIR-JP corpus does not reflect such strong usage. As shown in figure 1, although the NP-based pattern was still the largest structural type by Japanese writers (40% of bundle types), the overall number of bundles of this type (32.16% of bundle tokens) was smaller than VP-based bundles (42.24% of bundle tokens). This suggests that the Japanese professionals do not use noun phrases as consistently as their English counterparts. Instead, they demonstrate a reliance on

VP-based bundles when writing marine investigation reports. This finding is somewhat surprising: as the use of complex noun phrases has been found prevalent in written registers, particularly in specialized language (Biber 2006; Biber and Clark 2002; Breeze 2013), we assumed that NP-based bundles are among the most common multi-word sequences in both corpora. Reasons for the common use of verb phrases by Japanese writers are not completely clear. One plausible explanation might be that the English writing proficiency of Japanese professionals has not reached the advanced level. Such interpretation is derived from the claim that both L1 and L2 writers usually adopt a clausal (VP) style of discourse at an early stage, where many types of complex phrasal embedding (NP and PP constructions) are not acquired naturally at this stage (Biber, Gray and, Poonpon 2011: 29). This issue therefore deserves to be further explored in future studies.

In the following sections, the differences between the two corpora in each structural category will be elaborated.

4.2.2 Comparison of NP-based structural category across two corpora

Close scrutiny of NP-based category in both corpora reveals that "NP with of-phrase fragment" pattern made up the majority of this structural type, as shown in Table 6, and the frame "*the* + Noun + *of the/a*" was used as the most productive frame under this structural pattern.

Table 6. Proportional distribution of subcategory within NP-based pattern

NP-based Pattern	MAIR-EN		MAIR-JP	
	Bundle types	Bundle tokens	Bundle types	Bundle tokens
1) NP with of-phrase fragment	63%	70%	56%	60%
2) NP with other post-modifier fragment	14%	8%	24%	21%
3) Other NP expressions	23%	22%	20%	19%

By further examining the use of the frame "*the* + Noun + *of the/a*" in each corpus, significant difference is found between the two groups of writers. That is, compared with English writers, the Japanese professionals not only used this

frame relatively more frequently, but also employed a wider range of variants/nouns to fill in the frame.

This finding is illustrated in Table 7 below.

Table 7. Variants of the frame '*the* + Noun + *of* + *the/a*' across two corpora

Corpus	Variants in '*the* + Noun + *of* + *the/a*'	Total	
		type	Token (per million words)
MAIR-EN	**time**(443), requirements(111), purpose(84) use(70), **day**(64), **top**(64), operation(49), **vicinity**(46), safety(42), officer (41)	10	1014
MAIR-JP	**time**(613), occurrence(421), **vicinity**(228), statements(194), site(146), cause(136), **day**(116), statement(114), center(102), surface(86), bottom(82), end(79), situation(75), direction(68), crew(68), list(55), height(50), contents(50), location(46), inside(46), position(45), master(45), weight(45), results(42), middle(41), **top**(41)	26	3035*

Note: *: p <0.001, log likelihood=1203.56; The variants appearing in both corpora are indicated in bold; The token number of each variant is provided in brackets

As can be seen in the Table, the normalized frequency of the frame in the MAIR-JP corpus is 3,035 per million words (pmw), while it occurs only 1,014 times (pmw) in the MAIR-EN corpus. The log-likelihood test comparing these two numbers indicates that this frame is statistically overused by Japanese reporters. It is also noticeable from Table 7 that the types of variants employed by Japanese writers to fit in the slot were more than double that of by English counterparts (i.e. 26 types in MAIR-JP versus 10 types in MAIR-EN). Thus it can be inferred that the Japanese writers use this frame more flexibly. Moreover, there are only four variants shared by both corpora (i.e. *time, day, vicinity* and *top*), which signifies that these two groups of writers used this frame very differently from each other. Indeed, by conducting semantic tagging of each variant using UCREL Semantic Analysis System (USAS) (available at http://ucrel.lancs.ac.uk/usas/), we find that most nouns appearing in the slot in

MAIR-EN corpus fall into 2 major semantic fields, which are general/abstract terms (i.e. *operation, use, safety*) and psychological actions (i.e. *requirements, purpose*). However, Japanese professionals displayed a tendency to collocate the frame with the variants belonging to other semantic domains, such as location (i.e. *top, middle, position, inside, surface, direction, center, bottom, site vicinity, location*) and measurement (i.e. *height, weight, list*).

4.2.3 Comparison of PP-based structural category across two corpora

In terms of PP-based bundles, the study corpora still exhibit many differences between each other, as shown in Table 8.

Table 8. Proportional distribution of subcategory within PP–based pattern

PP-based Pattern	MAIR-EN		MAIR-JP	
	Bundle types	Bundle tokens	Bundle types	Bundle tokens
1) Prepositional phrase with embedded *of*-phrase fragment	43%	52%	27%	36%
2) Other prepositional phrase fragment	57%	48%	73%	64%

In the MAIR-EN corpus, two PP-based patterns occurs with similar proportion (around 50%) meaning English writers worked flexibly in constructing PP-based bundles. “Other prepositional phrase fragment” makes up the majority of this structural type in the MAIR-JP corpus (i.e. bundle types and tokens account for 73% and 64% respectively). Detailed investigation into this pattern in MAIR-JP corpus indicates that more than one third of lexical bundles within this subcategory start with preposition of, which is used the most frequently by Japanese professionals. Examples include of the accident site, of the crew members, of the port side, etc. The strong preference for this frame implies that Japanese writers tend to detail the information when writing MAIR, because prepositional bundles starting with of are often used to specify possessions (Biber et al. 1999). Such examples found in the MAIR-JP corpus include the possessions belong to any specific vessel (i.e. *of ship* A *and*); part of the ship (i.e. *of the hatch covers*); any accidents and investigation (i.e. *of the accident occurrence*),

even the people involved in (i.e. *of the chief officer*), etc.

4.2.4 Comparison of VP – based structural category across two corpora

4.2.4.1 Comparison of VP with *that*-clause pattern across two corpora

Close examination of all the VP-based bundles in each corpus indicated that the "VP with that clause" and "VP with active verb" (categories 9 and 13) patterns merited special attention. When constructing lexical bundles with the that clause, the two groups of writers adopt different strategies to control the clause. As for lexical bundles containing an active verb, the common occurrences of this type of bundles in MAIR is distinguished from the convention of academic prose. Therefore, in this subsection, the characteristics these two patterns exhibit in both corpora are presented in detail.

Lexical bundles incorporating clause fragment constitute the largest proportion of all VP-based bundles in the MAIR-JP corpus (72% of bundles types and 53% of bundle tokens). By contrast, these only make up a small percentage in the MAIR-EN corpus (10% of bundle types and 8.8% of bundle tokens). Hence it can be argued that Japanese writers favor *that*-clause structure much more than English writers. Interestingly, these two groups of writers are also found to use this pattern differently. To be specific, most of the *that*-clauses produced by English writers are controlled by main verbs in active voice, such as *warned the pilot that*, *stated that the vessel*, etc., while in the MAIR-JP corpus, except for *thought that it was*, all lexical bundles of this type belong to the pattern of adjective + *that*-clause. Even more remarkably, the adjectives used for controlling each *that*-clause convey certainty/uncertainty. Such examples include *is probable that it*, *is somewhat likely that*, *is highly probable that*, etc. Since different kinds of *that*-clauses serve different functions (Biber and Conrad., 2009), it can be inferred that for Japanese writers, the pattern of VP with that-clause fragments is a straightforward and perhaps more accessible way to express their stance towards the information. The following sentences extracted from the MAIR-JP corpus illustrate this point.

- It is *probable that Master* B was in good physical health when the accident happened.
- It is also *considered somewhat likely that* the foreman measured the O_2 concentration by himself from around 07:50 to around 08:05.

4.2.4.2 Comparison of VP with active verb pattern across two corpora

The presence of VP-based bundles with active verbs stands out in both corpora, which deserves further investigation. As shown in Table 5, the pattern of VP with active verbs is the most frequently used VP-based structure in the MAIR-EN corpus and it ranks 4th in the MAIR-JP corpus. This result appears some what surprising to us because passive voice is widely used in formal writings such as official reports or academic papers, in which actions themselves are often considered more significant than the agents of the actions (Oxford Dictionaries 2017). However, it is this point that precisely reflects the distinguished characteristics of the MAIR genre. It can therefore be concluded that the communicative purpose of the MAIR genre does not simply lie in reporting the accidents. Rather, it highlights the information about what or who caused or performed the activity, as seen in the following examples.

- About 0300, the chief officer *arrived on the bridge* to take over the navigational watch.(MAIR-ENcorpus)
- The purpose of the analysis is to determine the contributory causes and circumstances of the accident as a basis for *making recommendations to prevent* similar accidents occurring in the future. (MAIR-EN corpus)
- The master said that, because of the severe weather *leading up to* the accident, he had last slept on Tuesday morning, more than 48 hours before the accident. (MAIR-EN corpus)
- Master A *put the helm to* port in a step by step manner: first 10° and then 20°. (MAIR-JP corpus)
- At about 0739, while Vessel B was turning to port, the starboard center hull of *Vessel B collided with* the bow of Vessel A. (MAIR-JP corpus)
- At about 0740, while *Vessel A was proceeding* toward the direction of about 033° at about 7.9 knot. (MAIR-JP corpus)

4.3 Comparison of functional types of lexical bundles across two corpora

4.3.1 Comparison of distribution of functional types across two corpora

Lexical bundles used by English writers versus Japanese professionals were compared for their typical discourse functions based on the classification developed by Biber, Conrad, and Cortes (2004). A category of "subject-specific bundles" is added to the framework in order to classify the lexical bundles which are relating to the topic of the texts but their functions had not been identified by the established taxonomy. The additional category was then subdivided into four groups, namely, people, institutions, vessels, and equipment, on the basis of the entities they referred to. Table 9 provides a comprehensive classification of discourse functions, in which four main categories with 15 subcategories are involved. Similar to the comparison of structural distribution across two corpora discussed in the previous subsection 4.2.1, the statistical differences in the use of each functional type by two groups of writers were also ascertained by a log-likelihood test.

The results are presented in Table 9 as well.

Table 9. Functional distribution of lexical bundles across two MAIR corpora

Category	Structural patterns	Lexical bundle tokens (raw/normalized frequency per million words)	
		MAIR-EN	MAIR-JP
Stance bundles	1) Epistemic stance (*it is possible that*)	560 (302)	8600 (11721)*
	2) Attitudinal/modality stance (*it is desirable that*)	295 (159)	161 (219)*
Discourse organizers	3) Topic introduction/focus(*it is as follows*)	497 (268)	533 (726)*
	4) Topic elaboration/clarification(*on the other hand*)	-	114 (155)*
	5) Identification/focus (*course of the events*)	4570 (2467)	8550 (11653)*
	6) Tangible framing attributes (*at a speed of*)	105 (57)	2322 (3165)*
Referential bundles	7) Intangible framing attributes (*as a result of*)	3355 (1811)	**1378 (1878)**
	8) Place reference (*on both sides of*)	2145 (1157)	4811 (6557)*
	9) Time reference (*on the day of*)	3295 (1779)	2368 (3227)*
	10) Text Deictic (*as shown in the*)	-	56 (76)*

Subject-specific bundles	11) people (*the officer of the*)	819 (442)	1362 (1856)*
	12) Regulations (*port marine safety code*)	814 (439)	189 (258)*
	13) Vessels (*that the vessel was*)	372 (201)	1514 (2063)*
	14) Institutions (*maritime and coastguard agency*)	1520 (820)	1019 (1389)*
	15) Equipments (*the main engine to*)	74 (40)	428 (583)*
Total		18421(9942)	33405 (45448)*

Note: *: $p < 0.001$; the number marked in bold: $p = 0.257$, log likelihood = 1.287

Table 10. Functional distribution of lexical bundles across two MAIR corpora

Corpus	Variants in 'the + Noun + of + the/a'	Total	
		type	Token (per million words)
MAIR-EN	time(443), requirements(111), purpose(84) use(70), day(64), top(64), operation(49), vicinity(46), safety(42), officer (41)	10	1014
MAIR-JP	time(613), occurrence(421), vicinity(228), statements(194), site(146), cause(136), day(116), statement(114), center(102), surface(86), bottom(82), end(79), situation(75), direction(68), crew(68), list(55), height(50), contents(50), location(46), inside(46), position(45), master(45), weight(45), results(42), middle(41), top(41)	26	3035*

Note: *: p <0.001, log likelihood=1203.56; the variants appearing in both corpora are indicated in bold; the token number of each variant is provided in brackets

As shown in Table 10, all functional types are statistically overused by Japanese reporters, except for the discourse function of intangible framing attributes. It can therefore be concluded that the overuse tendency of Japanese writers is still obvious in the functional distribution of lexical bundles.

When considering which functional categories contribute the most to the difference between the corpora, we calculated the value of standardized residuals (R) in a chi-square contingency table, as shown in Table 11.

Table 11. Standardized residuals in a chi-square contingency table for functional distribution (tokens)

χ^2=24.623, *df*=3,*p*<0.001, Cramer's V=0.234		Stance bundles	Discourse organizers	Referential bundles	Subject-specific bundles
MAIR-EN	Observed Count	4.6	27.3	48.0	19.4
	Expected Count	22.2	27.4	35.3	14.5
	R	**-3.7**	0	**2.1**	1.3
MAIR-JP	Observed Count	119.4	125.3	149.0	61.5
	Expected Count	101.8	125.3	161.7	65.7
	R	1.7	0	-1.0	-0.6

Note: the token numbers used here are normalized frequency; the cells with the absolute R value greater than 1.96 are in bold.

As Table 11 illustrates, there is a significant difference in the functional distribution of bundle tokens between two corpora at the 0.05 level (χ^2= 24.623, p<0.001), which is in accordance with the result obtained by a log-likelihood test. More importantly, it is found that R values for the cells of stance bundles and referential bundles in the MAIR-EN corpus are -3.7 and 2.1. Both are greater than 1.96 in absolute value suggesting that these two functional types make a statistically significant contribution to the difference.

Based on the above findings, further investigations were carried out only within these two categories rather than all three functions. The results will be discussed in the following subsections.

4.3.2 Comparison of stance bundles across two corpora

Stance bundles were commonly used by both groups of writers. They are extremely common in the MAIR-JP corpus, ranking as the largest category among all functions. This means that Japanese writers rely heavily on stance bundles when writing MAIR. When looking at the distribution characteristics within the category of stance bundles, it is noticeable that epistemic bundles were preferred by both groups of writers to express stance. Such evidence can be found from the proportions that epistemic bundles take in both corpora, as seen in Table 12.

Table 12. Proportional distribution of subcategory within stance bundles

Stance bundles	MAIR-EN		MAIR-JP	
	Bundle types	Bundle tokens	Bundle types	Bundle tokens
1) Epistemic stance	67%	76%	94%	99%
2) Attitudinal/modality stance	23%	24%	6%	1%

As shown in Table 12, epistemic bundles make up the majority of stance bundles in the MAIR-EN corpus, having percentages at 67% of bundle types and 76% of bundle tokens respectively. More strikingly, it holds an absolutely dominant position among stance bundles in the MAIR-JP corpus, accounting for 94% of bundle types and 99% of bundle tokens. Given that epistemic bundles are used to "comment on the knowledge status of the information in the following proposition" (Biber, Conrad, and Cortes 2004), the preference for such bundles can be understood as a preferred form that both groups adopted to assess accidents based on investigation results. Beyond that, all epistemic bundles in two corpora were also found to be impersonal, such as it is likely that, it is possible that, it is considered probable that, it is considered somewhat likely, etc, which indicates that both groups of writers laid emphasis on minimizing the imposition of their opinions when expressing assessments.

Another interesting finding concerning epistemic bundles is that none of the expressions are shared by the two groups of writers. For instance, lexical bundles embedded with probable were commonly employed by Japanese writers to express their tentative stance (e.g. *it is probable that*, *it is considered probable*, etc.). On the other hand, similar expressions do not occur in the MAIR-EN corpus. Instead, epistemic bundles containing possible were frequently used by English writers to mitigate the proposition (e.g. *it is possible that*).

Even though some lexical bundles incorporate the same word, the two groups of writers constructed them quite differently. Examples can be found from the usage of lexical bundles containing likely. In the MAIR-EN corpus, one bundle type it is likely that was employed by the English reporters to hedge their statements. However, among all the likely bundles used by Japanese writers, the adverb somewhat always co-occurred with likely as a pre-modifier to express a low degree of certainty about the accidents being investigated, as seen

in the following examples.

- *It is somewhat likely* that the stanchion on top of the bulwarks on the starboard side was placed to prevent the piled up pearl nets from sliding over the edge of the vessel. (MAIR-JP corpus)
- It is therefore *considered somewhat likely that* Pilot A2 kept on accelerating Vessel A because she was supposed to go ahead of Vessel C. (MAIR-JP corpus)
- *It is considered somewhat likely* that Master B did not put the rudder to starboard because there was a Light buoy on the starboard side of Vessel B and the water outside the passage was not deep enough. (MAIR-JP corpus)

The presence of this adverb in the use of epistemic bundles implies that Japanese writers are more cautious when drawing inferences on the basis of the investigation results. In other words, English writers demonstrate better control expressing the degree of doubt and certainty.

4.3.3 Comparison of referential bundles across two corpora

As discussed above, another noticeable functional difference between MAIR-EN and MAIR-JP corpus appears in the category of referential bundles. Within the MAIR-JP corpus, a comparison across all subcategories indicated that lexical bundles functioning as tangible frames, time and place references were used much more than other discourse functions, as seen in figure 2 below. As the primary functions of these types of referential bundles are to make direct reference to physical entities, time and places or to single out the natural attributes of the entities (Biber, Conrad, and Cortes 2004), it can be inferred that Japanese writers paid more attention to describing and detailing the information in their MAIR writing. For instance, Japanese writers employed a wide range of tangible frames either to specify the attributes of the vessels (e.g. *at a speed of, at an angle of, the list of, the weight of the, the position of the, with a heading of,* etc) or to describe the accident occurrences (e.g. *the center of the, the surface of the, at a distance of,* etc). The MAIR-JP corpus also contains a number of tangible framing bundles referring to the statements of the individuals under investigation, as the following examples demonstrate.

- According to *the statement of the* staff of Company B, the person who made the VHF calls from Vessel B to Vessel A at about 02:05:26 and 02:05:37 was identified as Officer B on board Vessel B. (MAIR-JP corpus)
- According the oral statements and *the written reply to* the questionnaire from the person-in-charge of the Charterer, the situation is as follows. (MAIR-JP corpus)

In the MAIR-EN corpus, although there is a group of lexical bundles referring to time and places, only one lexical bundle serves for tangible framing attribute, which is *at the speed of*, occurring 58 times per million words. Therefore, we can deduce that, English writers do not rely on this discourse function as much In contrast, lexical bundles functioning as intangible frames make up a significant proportion of all referential bundles (44% of bundle types and 38% of bundle tokens) in the MAIR-En corpus, as shown in Figure 2.

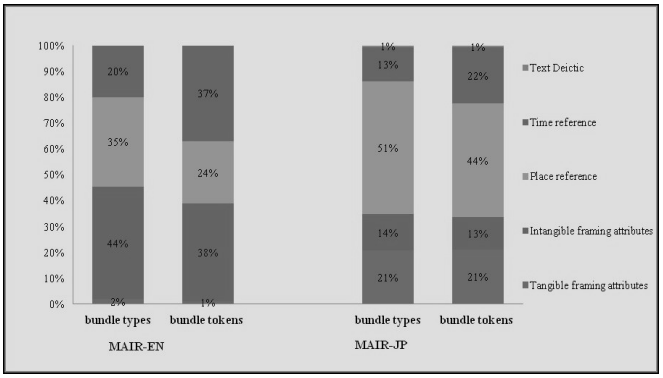

Figure 2. The proportional distribution of referential bundles across six subcategories in two MAIR corpora (type and token)

Apart from that, a close look at the data under this subcategory shows that most bundle types are used to specify the cause or result of the accident or the condition of the occurrences. See the following examples:

- There were 131 passengers and crew on board the Marchioness, 51 of whom died *as a result of* the accident. (MAIR-EN corpus)

- He had not had any assessment of his performance as master at sea during the 3 years *leading up to the* accident. (MAIR-EN corpus)
- *Contributing to the accident* was the first pilot's fatigue, caused by his untreated obstructive sleep apnea and his work schedule, which did not permit adequate sleep. (MAIR-EN corpus)
- *In the case of* crew training deficiencies, an extended time limit is given to obtain the necessary training. (MAIR-EN corpus)
- Due to *the circumstances of this* accident, and the consequent absence of survivors and material evidence, its causes remain a matter of some speculation. (MAIR-EN corpus)

As Biber, Conrad, and Cortes (2004) illustrated, one discourse function of the referential bundles is to identify abstract characteristics. This function is primarily used for establishing logical relationships in the text. Thus it turns out that English reporters tend to focus on analyzing the accident occurrences when writing MAIR. In particular, they place emphasis on determining the relationship between any specific causes/conditions and results, which may be considered one of the major distinctions from the Japanese reporters' writing.

5. Conclusion

This study attempted to understand the lexical bundles constructed by writers from different linguistic backgrounds in an ESP context. Through comparison of lexical bundles used by L1-English versus L1-Japanese professionals in the MAIR, differences between these two groups of writers have been clearly displayed. In general, compared with English reporters, Japanese professionals employ a considerably wider range of four-word bundles, exhibit an overuse tendency in almost all structural patterns and functional types, and adopt different strategies to construct lexical bundles and to fulfill discourse functions. For instance, English reporters are found to use NP-based bundles to a large extent, but Japanese writers demonstrate a reliance on VP-based bundles when writing accident investigation reports; It is also found that native-English experts construct PP-based bundles flexibly and do not much favor the

that-clause. Japanese professionals, however, prefer PP-based bundles starting with preposition of and *that*-clause fragment, controlled by adjectives which convey certainty/uncertainty. In terms of stance bundles, these two groups of writers use distinct expressions, even though the lexical bundles they employ include the same epistemic adjectives. Moreover, adverbs are always embedded into likely-bundles by the Japanese writers to express a low degree of certainty but English reporters do not behave the same way. The use of referential bundles is also different between two groups of writers, where Japanese experts use a great number of lexical bundles functioning as tangible frames, time and place references. English professionals, by contrast, are much dependent on the use of referential bundles, which have intangible framing attributes. The differences mentioned above imply that Japanese writers pay more attention to detailing the information in their MAIR writing, and are cautious about drawing inferences. Conversely, English reporters tend to analyze the accident occurrences, with a special emphasis on establishing the cause/effect relationship. Beyond that, they demonstrate better control expressing the degree of doubt and certainty than Japanese writers.

Notably, two groups of writers also display similarities in the use of some types of lexical bundles, which are able to reflect the special characteristics of MAIR genre. First, active verbs are commonly used by two groups of writers, indicating that the communicative purpose of MAIR genre does not simply lie in reporting the accidents. Rather, it focuses on highlighting the information about what or who caused or performed the activity. Meanwhile, the wide use of epistemic bundles in both corpora reflects that accident evaluation is also a significant part of MAIR genre. When expressing assessments, both groups of writers display a tendency to mitigate the imposition of their opinions.

These findings not only provide insights into the nature of how lexical bundles are used by authors from different linguistic backgrounds in the genre of MAIR but also are likely to raise the readers' awareness of the existing varieties within the genre and offer them easy access to MAIR.

Finally, it has to be admitted that there are some unavoidable limitations of this study, one of which is that the reasons for differences between Japanese and English professionals are not delved in depth. The study might have been more fruitful if it had discussed the conceptual and cultural motivations behind these

discrepancies, since any forms of language and linguistic choices are believed to bear its cultural implication.

References

Ädel, Annelie and Britt Erman. 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.

Bernardini, Silvia, Adriano Ferraresi, and Federico Gaspari. 2010. Institutional academic English in the European context: A web-as-corpus approach to comparing native and non-native language. In Angeles Linde Lopez and Rosalia Crespo Jimenez (eds.), *Professional English in the European context: The EHEA challenge*, 27-53. Bern: Peter Lang.

Biber, Douglas. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5: 257-269.

Biber, Douglas. 2003. Variation among university spoken and written registers: A new multi-dimensional analysis. In Charles Meyer and Pepi Leistyna (eds.), *Corpus analysis: Language structure and language use*, 47-70. Amsterdam: Rodopi.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, Douglas and Bethany Gray. 2011. Grammatical change in the noun phrase: the influence of written language use. *English Language and Linguistics* 15(2): 223-250.

Biber, Douglas and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26: 263-286.

Biber, Douglas and Susan Conrad. 2009. *Register, genre and style*. Cambridge: Cambridge University Press.

Biber, Douglas and Victoria Clark. 2002. Historical shifts in modification patterns with complex noun phrase structures: How long can you go without a verb? In Teresa Fanego, Maria Jose Lopez-Couso, and Javier Perez-Guerra (eds.), *English historical syntax and morphology*, 43-66. Amsterdam: John Benjamins.

Biber, Douglas, Bethany Gray, and Kornwipa Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45(1): 5-35.

Biber, Douglas, Susan Conrad, and Viviana Cortes. 2004. If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371-405.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education.

Breeze, Ruth. 2013. Lexical bundles across four legal genres. *International Journal of Corpus Linguistics* 18(2): 229-253.

Chen, Yu-Hua and Paul Baker. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology* 14: 30-49.

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

Daille, Béatrice. 1995. *Combined approach for terminology extraction: lexical statistics and linguistic filtering*. Lancaster: University Center for Computer Corpus Research on Language.

De Cock, Sylvie. 2000. Repetitive phrasal chunkiness and advanced EFL speech and writing. In Christian Mair and Marianne Hundt (eds.), *Corpus Linguistics and Linguistic Theory*, 51-68. Amsterdam: Rodopi.

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

Erman, Britt. 2009. Formulaic language from a learner perspective: What the learner needs to know. In Roberta Corrigan, Edith A. Moravcsik, Hamid Quali, and Kathleen M. Wheatley (eds.), *Formulaic language*, 27-50. Amsterdam: John Benjamins.

Gaspari, Federico. 2013. A phraseological comparison of international news agency reports published online: Lexical bundles in the English language output of ANSA, Adnkronos, Reuters and UPI. In Magnus Huber and Joybrato Mukherjee (eds.) *Corpus Linguistics and Variation in English: Focus on Non-Native Englishes. Proceedings of ICAME 31. VARIENG : Studies in Variation, Contacts and Change in English* 13. University of Helsinki: Research Unit for Variation, Contacts, and Change in English. Retrieved from http://www.helsinki.fi/varieng/series/volumes/13/gaspari/ on 15 May, 2017.

Goźdź-Roszkowski, Stanislaw. 2011. *Patterns of linguistic variation in American legal English: A corpus-based study*. Frankfurt am Main: Peter Lang.

Grabowski, Lukasz. 2015. Phrase frames in English pharamaceutical discourse: A corpus-driven study of intra-disciplinary register variation. *Research in Language* 13(3): 266-291.

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In Anthony P. Cowie (ed.), *Phraseology: Theory, analysis and applications*, 145-160. Oxford: Clarendon Press.

Howarth, Peter. 1998. Phraseology and second language proficiency. *Applied Linguistics* 19(1): 24-44.

Hyland, Ken. 2008a. Academic clusters: text patterning in published and postgraduate writing. *Journal of Applied Linguistics* 18(1): 41-62.

Hyland, Ken. 2008b. As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4-21.

IBM. 2012. *IBM SPSS Statistics for Windows (Version 21.0)*. [Computer Software] Armonk, New York: IBM.

Jablonkai, Réka. 2010. English in the context of European integration: A corpus-driven

analysis of lexical bundles in English EU documents. *English for Specific Purposes* 29: 253-267.

Jhang, Se-Eun and Sung-Min Lee. 2013. Clusters and key clusters in the Maritime English Corpus. *Journal of Language Sciences* 20(4): 199-219.

Jukneviciene, Rita. 2009. Lexical bundles in learner language: Lithuanian learners vs. native speakers. *Kalbotyra* 61(3): 61-71.

Koester, Almut. 2006. *Investigating workplace discourse*. Abingdon, Oxford: Routledge.

Lamart, Stephanie. 2013. *Standardized residuals in statistics: What are they?* Retrieved from http://www.statisticshowto.com/probability-and-statistics/statistics-definitions/ on 25 May, 2017.

Lewis, Margareta. 2009. *The idiom principle in L2 English: Assessing elusive formulaic sequences as indicators of idiomaticity, fluency, and proficiency*. Saarbrücken: VDM Verlag.

Lu, Wenyu, Sung-Min Lee, and Se-Eun Jhang. 2017. Keyness in maritime institutional law texts. *Linguistic Research* 34(1): 51-76.

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus based language studies: an advanced resource book*. Abingdon, Oxford: Routledge.

Miller, Carolyn Rae. 1984. Genre as social action. *Quarterly Journal of Speech* 70: 151-176.

Nekrasova, M. Tatiana. 2009. English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning* 59: 647-686.

Oxford Dictionaries. 2017. Retrieved from https://en.oxforddictionaries.com/ on 10 May, 2017.

Pan, Fan, Randi Reppen, and Douglas Biber. 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.

Römer, Ute. 2009. The inseparability of lexis and grammar. *Annual Review of Cognitive Linguistics* 7: 141-163.

Schmitt, Norbert and Ronald Carter. 2004. Formulaic sequences in action: An introduction. In Norbert Schmitt (ed.), *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins.

Scott, Mike. 2014. *Wordsmith tools (Version 6.0)*. [Computer Software] Liverpool: Lexical Analysis Software.

Scott, Mike and Christopher Tribble. 2006. *Textual patterns*. Amsterdam: John Benjamins.

Simpson, Vlach, Rita and Nick C. Ellis. 2010. An academic formulas list: new methods in phraseology. *Research Applied Linguistics* 31(4): 487–512.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Se-Eun Jhang
Professor
Department of English Language and Literature
Korea Maritime and Ocean University
727 Taejong-ro, Yeongdo-gu, Busan 49112, Korea
E-mail: jhang@kmou.ac.kr

Sungkuk Kim
Instructor
Institute of International Trade
Sungkyunkwan University
25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea
E-mail: lloyds@skku.edu

Yilian Qi
Lecturer
English Department
Dalian Maritime University
No.1 Linghai Rd, Ganjingzi District, Dalian, Liaoning Province116000, China
PhD Student
Department of English Language and Literature
Korea Maritime and Ocean University
727 Taejong-ro, Yeongdo-gu, Busan 49112, Korea
E-mail: yilian_qchi@hotmail.com

Received: 2017. 09. 30.
Revised: 2018. 06. 28.
Accepted: 2018. 06. 28.