Linguistic Research 34(1), 51-76
DOI: 10.17250/khisli.34.1.201703.002

Keyness in maritime institutional law texts*

Wenyu Lu[a b] · Sung-Min Lee[b] · Se-Eun Jhang[b]**
(Dalian Maritime University[a] · Korea Maritime and Ocean University[b])

Lu, Wenyu, Sung-Min Lee, and Se-Eun Jhang. 2017. Keyness in maritime institutional law texts. *Linguistic Research* 34(1), 51-76. This study describes some characteristics of maritime institutional legal texts in terms of corpus methodology. We self-built two study corpora: a public maritime institutional corpus and a private maritime institutional corpus. The differences between the two corpora can be distinguished by identifying typical linguistic features from the keyness aspects of key words, key clusters, and key semantic domains. Specific words and phrases in complementary distribution are offered to distinguish public maritime legal characteristics from private maritime legal characteristics by comparing the self-built specialized corpora with the more general British National Corpus (BNC informative genre). Linguistic features are discussed from the view of keyness, thus enabling non-legal practitioners as well as non-specialist readers to discover and describe underlying parameters that best depict the differences between legal registers or genres. **(Dalian Maritime University · Korea Maritime and Ocean University)**

Keywords maritime institutional texts, public law, private law, keyness, complementary distribution

1. Introduction

In corpus linguistics, specialized English corpora can assist in recognition of language used in specific areas which is very different from general English. According to Johns (2013: 5), the origin of English for Specific Purposes (ESP) corpus-based research can be dated back to the 1960s when the central focus of ESP research was English for science and technology (EST) in academic contexts. The research at that time was mostly descriptive, involving a few statistical grammatical counts within written discourse. In the 1990s, along with the development of corpus linguistics, researchers started to pay special attention to different subfield,

* The authors would like to thank the anonymous reviewers for their valuable comments on the earlier version of this article.
** Corresponding Author

particularly specific written academic registers. For example, as a particular ESP genre, legal English has drawn many researchers' interests. Bhatia (1993) analyzed syntactic features of legislative texts including sentence length, nominalization, complex prepositional phrases, binomial and multinomial expressions, initial case descriptions, qualifications in legislative provisions and syntactic discontinuities. Maley (1994) provided a taxonomy of legal genres and emphasized that the development of a special legal language represented "a predictable process and pattern of functional specialization" (Maley 1994: 11). Iber (2001) described a task-based course incorporating the use of concordances for legal essay writing. Hafner and Candlin (2007) evaluated students' use of a simple online concordance tool in completing legal writing tasks such as drafting legal opinions of court pleadings. Some scholars such as Šarčević (1997) and Asensio (2003) discussed the methods of translating official legal documents. Goźdź-Roszkowski (2011) focused on the patterns of linguistic variation in American legal English. Moreover, suicides' posthumous papers, witnesses' testimony, policemen's hearing records, etc. were also collected to build specialized corpora, such as the Communicated Threat Assessment Resource Corpus (CTARC), Corpus of Supreme Court Opinions (COSCO), and Corpus National University of Singapore Short Message Service Corpus (NUSSMS).

Under the definition of maritime English,[1] English for maritime law is different from general English through its own particular characteristics. Since ancient times until the emergence of modern national states, the law governing maritime commerce had been largely uniform in the western world. From the fifteenth to the seventeenth century, with the coming of great maritime era, western countries started to develop their navigation technology. The corresponding commerce across countries started to boom. Entering the eighteenth and nineteenth centuries, an internationally-accepted "law of the sea" was badly in need to solve the disputes in the maritime field across involving various countries.

Maritime law (also known as admiralty law) includes institutional texts (i.e., international conventions, rules and regulations, agreements, etc.), maritime legal

---

1 According to Bocanegra-Valle (2013: 3579-3580), Maritime English refers to the English language used by seafarers both at sea and in port and by individuals working in the shipping and shipbuilding industry. It subsumes five different sub-varieties according to the specific purpose they serve within the maritime context: English for navigation and maritime communications, English for maritime commerce, English for maritime law, English for marine engineering, and English for shipbuilding.

proceedings and their relative documents (reports of lawsuits, judgments, court rulings and decisions, case surveys, transcripts of court hearings, out-of-court settlements, arbitration awards, company agreements, mergers, etc.) issued by maritime courts and other legislative bodies. Specifically, maritime institutional texts are legal regulations governing maritime shipping and relevant activities, the use of the sea, the exploitation of its resources and the protection of the marine environments. These are compiled from three fields: national maritime law, international public maritime law and international private maritime law. National maritime laws are only accepted by the countries themselves. International public maritime law generally concerns matters related to the distribution and exercise of power by public authorities and the legal relations between the State (and its administration) and individuals, including the registration of vessels, safety of ships and safety of navigation, control of shipping operations, the movement of persons and goods in port, casualty investigations and some aspects of preservation and protection of the marine environments. International private maritime law is concerned with legal relationships between individuals or groups of individuals such as co-operatives and companies. Its primary purpose is the protection of individual interests such as the acquisition or transfer of the ownership of vessels, charter-parties, and bills of lading.

Both international public and private laws are effective and should be enforced among the signatory countries who sign the convention. However, there are no clear standards to sort existing institutional legal texts into international public or international private categories. Most of these determinations are made by maritime lawyers. For this reason, it is necessary to compile a Maritime Legal English Corpus (MLEC) which can offer authentic maritime legal language usage for ESP learners, seafarers and maritime lawyers. This study aims to distinguish a public maritime institutional corpus from a private maritime institutional corpus by identifying their typical linguistic features. The previous corpus-based approach to maritime legal languages is largely based on general description, discussing frequencies of linguistic features, distribution patterns and sentence complexity by looking at word lists and concordance lists. Hong and Jhang (2010) compiled a general Maritime English Corpus, but focused only on lexicon. Jhang and Lee (2013) included several important conventions into their corpus, but did not offer a detailed analysis of maritime institutional texts. Orts-Llopis (2009, 2014) focused more on the genre of

delegated legislation and tenancy agreements or leases and the mechanics that articulate particular form of contract as a genre rather than conventions and regulations. Most recently, Lee (2016) included some maritime institutional law texts in his 4 million words' corpus, but mainly discussed both collocation network analysis and keyword network analysis in order to identify general terms and specific terms respectively.

2. Data and methods

2.1 Study and reference corpora

As a specialized agency of the United Nations, the International Maritime Organization (IMO) is the global standard-setting authority for the safety, security, and environmental performance of international shipping. Its main role is to create a regulatory framework for the shipping industry that is fair and effective, universally adopted, and universally implemented. As an internationally important authority, IMO has promoted the adoption of about 30 conventions and protocols, nearly all of which are now in force. Conventions and protocols are binding legal instruments, and upon entry into force their requirements must be implemented by all countries which are party to them. Appendix A shows the list of conventions mentioned in the IMO's web pages[2]. These conventions have been downloaded and converted into plain text files to construct the study (target) corpus. The conventions can be generally categorized into three aspects: (a) conventions relating to maritime safety and security and ship/port interface, (b) conventions relating to prevention of marine pollution, and (c) conventions covering liability and compensation. After consulting several legal professionals, we classified categories (a) and (b) as international public maritime law, type (c) as international private maritime law. Considering the confidential properties of juridical judgments, these two new study corpora are principally comprised of maritime conventions. The informative genre of the written subcorpus in the British National Corpus Sampler[3] (hence, BNC Sampler - Written

---

2 Information of the conventions can be found in [http://www.imo.org/en/About/Conventions/ListOfConventions/ Pages/Default.aspx].

3 British National Corpus Sampler can be free downloaded form [http://ota.ox.ac.uk/desc/2551].

- Informative) includes pure science, applied science, social science, world affairs, commerce and finance, belief and thought, arts, and leisure, whose language style is quite similar to legal documents, and was therefore chosen to be a reference corpus of general English.[4]

General information of the two study corpora and the reference BNC corpus is summarized in Table 1 below.

Table 1. General information of the two study corpora and a reference corpus

Corpora	Tokens	Types	Standardized Type-Token Ratio (STTR)
Private Law	50,578	2,241	26.49
Public Law	325,392	7,673	28.00
BNC Sampler-Written-Informative	779,027	38,629	43.11

The standardized type-token ratio (STTR) was computed per 1,000 words as a word list goes through each text file, as this is an effective way to show a variety of vocabulary in the corpus. It can be seen from Table 1 that the value of STTR in general English (43.11) is much higher than in maritime legal English (26.49; 28.00). This can be explained by the fact that as a specialized genre, many of the word types used in maritime institutional texts are used repeatedly, whereas those in BNC Sampler-Written-Informative are used in a variety.

2.2 Tools and methods

WordSmith 6.0 (Scott 2014) was used to analyze key words and key clusters of maritime institutional texts. We also used a Wmatrix web interface program (cf. http://ucrel.lancs.ac.uk/wmatrix3.html) for the UCREL semantic analysis system (USAS) (Rayson *et al.* 2004; Rayson 2009) based on McArthur's (1981) Longman Lexicon of Contemporary English to analyze the key semantic domains. Key words were analyzed based on a multi-tier structure with 21 major semantic categories.

4 Goh (2011) explored what factors of the reference corpus influenced the results of keyword calculation in a significant way, and pointed out that genre was one of more important factors to consider than other factors such as corpus size and varietal difference when choosing a reference corpus.

Each major category was further fine-grained into several subcategories. There are total of 113 subcategories: these subcategories can be further fine-grained into subgroups.

The term "key words" is widely used across various fields of study. From a linguistic point of view, it contributes to the long "search for units of meaning" (Sinclair 1996). From a sociological point of view, it is part of "a vocabulary of culture and society" (Williams 1976, 1983). In corpus linguistics, its concept is explicitly defined by Scott (1997: 236) as words which co-occur with unusual frequency from a study corpus compared with a reference corpus. Scott's approach to key words provides an empirical discovery method, based on frequency and distribution to discover the underlying features of global texts according to different kinds of keyness calculation methods. According to Scott (2015), keyness is a term used in corpus linguistics to describe the quality a word or phrase has of being "key" in its context. It is a textual feature, not a linguistic feature because a word may have keyness in a certain textual context but may not have keyness in other contexts.

In order to find the most effective "key" terms for a study corpus, we set the minimum frequency as 3 and the minimum percentage of texts as 5%. A minimum frequency helps to eliminate words or clusters which are unusual but infrequent, so as to reduce spurious hits. The minimum percentage of texts allows researchers to ignore words which are not found in many texts. Therefore, in this research, a word which occurs in the study corpus at least 3 times and occurs over at least 5% of those texts was considered a candidate key word. The candidate key word was then frequency-tested against a reference corpus to assess the statistical probability as computed by an appropriate algorithm. At last, all the key words were sorted by keyness, where positive key words are sorted over negative ones. According to Scott (2015), a word with positive keyness occurs more often than would be expected by chance in comparison with the reference corpus. A word which is negatively key occurs less often than would be expected by chance in comparison with the reference corpus.

Keyness is a value calculated by standard statistical tests like log-likelihood or chi-square. In the present study, a log-likelihood test (Dunning 1993) was chosen since it gives better results of key words and key clusters, particularly when contrasting long texts or a whole genre against a reference corpus (Scott 2015). Log-likelihood is calculated by first constructing a contingency table to calculate the expected values (E) by observed values (O). Then the values are calculated through

the following formula.

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \quad LL = 2*((a*\log(\frac{a}{E_1})) + (b*\log(\frac{b}{E_2})))$$

The *p*-value ranges from 0 to 1. A value of 0.05 would give a 5% danger of being wrong in claiming a relationship. In social sciences, a 5% risk is usually considered acceptable. In the case of key words and key clusters, where the notion of risk is less important than that of selectivity, we set a comparatively low *p*-value threshold such as 0.000001 (one in 1 million) (1E-6 in scientific notation) so as to obtain fewer key words. The higher the log-likelihood value, the more significant is the difference between two frequency scores (99.9999th percentile; 0.0001% level; $p < 0.000001$; critical value $\approx$ 24). The accuracy rate for the semantic tagger is 91%-92% (Rayson *et al.* 2004). After the completion of analysis by the software, we performed manual proof-checking before entering into serious discussions of the tagging results.

3. Results and discussion

3.1 Key words

3.1.1 Analysis of key words in complementary distribution

For the public law corpus, 1,392 key words were generated, among which 1,044 have positive keyness. For the private law corpus, 425 key words were generated among which 384 have positive keyness. Table 2 lists the top 20 and bottom 10 key words in the two study corpora.

Table 2. Key words lists of public / private law corpora

Public Law			Private Law		
N	**Key Word**	**Keyness**	**N**	**Key Word**	**Keyness**
1	SHALL	14816.86	1	SHALL	4573.55

2	SHIP	5724.00	2	CONVENTION	3737.65
3	REGULATION	4783.74	3	ARTICLE	3679.01
4	OR	4411.59	4	OR	2290.83
5	CONVENTION	4145.38	5	PARAGRAPH	2192.87
6	SHIPS	3052.31	6	STATE	1718.17
7	CERTIFICATE	2142.92	7	FUND	1498.11
8	OF	2093.87	8	DAMAGE	1448.94
9	ADMINISTRATION	2056.72	9	LIABILITY	1189.37
10	BE	1990.00	10	SHIP	1161.49
11	ORGANIZATION	1987.93	11	ANY	896.71
12	ARTICLE	1944.02	12	ACCORDANCE	839.61
13	PARAGRAPH	1830.41	13	OWNER	757.56
14	ANNEX	1807.34	14	PROTOCOL	670.65
15	DATE	1706.41	15	SUCH	638.47
16	CARGO	1596.78	16	COMPENSATION	634.74
17	ACCORDANCE	1562.21	17	ORGANIZATION	619.80
18	REQUIREMENTS	1506.04	18	DATE	617.46
19	SUCH	1503.18	19	STATES	612.74
20	PRESENT	1356.67	20	SECRETARY-GENERAL	608.09
...	...	...	...	...	...
1044	FEELING	23.93	384	CASES	24.14
1045	ON	-23.96	385	MAKE	-26.58
...	...	...	...	...	...
1383	CAN	-597.33	416	ARE	-136.98
1384	IT	-612.44	417	THEY	-154.57
1385	WOULD	-710.76	418	IT	-169.08
1386	WILL	-781.91	419	BUT	-172.94
1387	BUT	-944.72	420	HE	-182.11
1388	HIS	-1148.29	421	HAD	-208.42
1389	WERE	-1251.82	422	WILL	-243.19
1390	HE	-1317.37	423	AND	-253.42
1391	HAD	-1446.83	424	WERE	-256.53
1392	WAS	-2572.25	425	WAS	-411.99

Several interesting observations can be found here. First, although *regulation* and *administration* have very high ranking (Top 3, Top 9) in the public law corpus, they are not in the key words list of private law corpus at all; whereas *liability* and *compensation* are two words that have very high ranking (Top 9, Top 16) in the private law corpus, but they are not in the key words list of public law corpus at all. Complementary distribution of these two pairs of words in the key words lists can

help us to classify legal documents into public and private law. If the way in which we classify public and private corpora is straightforward, then in reverse, if a text has the word *administration*, for example, as its key word, we may confidently say that it is a convention in the range of public law. This complementary difference, therefore, offers a linguistic means to differentiate public and private law, without the help of legal experts. We will discuss the application of complementary distribution further in Section 6.

Second, the modal auxiliary verb *shall* is ranked highest in the both key words lists. This auxiliary exists in most legal documents, where *shall* is used to express the meaning of "compulsory orders".

Third, words like *of* and *or* appear in the top 10 key words list. In English, two nouns are often connected by *of* to denote ownership and two coordinate phrases or clauses are frequently held together by *or* to display options. This observation results in the comparatively longer length of sentence in legal documents, which makes comprehension of legal English rather more difficult by the general populace. Moreover, the high ranking of *or* instead of *and* indicates that maritime institutional texts favor options rather than coexistence.

At last, negative keyness shows underuse of a particular word when compared with a reference corpus. In maritime institutional legal texts, the auxiliary indicating past tense like *was, were* and *had* are unusually infrequent, proving that legal documents tend to use present tense to exhibit its objectivity and authority. Third person pronouns such as *he*, *they* and *it* are also underused. This is quite different from other genres, and may be explained by the fact that in maritime institutional legal texts, in order to precisely distribute the ideas, pronouns are underused to avoid ambiguity.

3.1.2 Analysis of modal verbs

The "central" English modals are usually considered to be *will, would, can, could, may, might, shall, should,* and *must*. In addition to these nine central modals, there is a small group of "marginal modals", *ought to, need to, used to* and *dare*, which can behave in some ways like modals and in other ways like main verbs (Quirk *et al.* 1985: 135). Extensive research has demonstrated that in legal documents, *shall* and *may* are in most frequent use (Hong and Jhang 2010: 978). As

for connotation, *shall* focuses on the "obligation" while *may* entitles the "rights".

These observations can also be found in maritime institutional texts, as displayed in Table 3 below.

Table 3. Modal auxiliary verbs in public / private law corpora

Public Law					Private Law				
N	**Key word**	**Keyness**	**Freq**	**%**	**N**	**Key word**	**Keyness**	**Freq**	**%**
1	SHALL	14816.86	6338	1.85	1	SHALL	4573.55	908	1.71
62	MAY	555.83	1262	0.37	61	MAY	263.34	264	0.50
1270	MIGHT	-85.19	31	0.00	398	MUST	-45.43	8	0.02
1287	SHOULD	-101.09	267	0.08	400	SHOULD	-52.04	23	0.04
1363	MUST	-302.07	33	0.00	405	WOULD	-65.79	36	0.07
1379	COULD	-531.60	23	0.00	526	COULD	-107.52	2	0.00
1385	CAN	-597.34	113	0.03	414	CAN	-121.87	15	0.03
1385	WOULD	-710.76	87	0.03	422	WILL	-243.19	12	0.02
1386	WILL	-781.91	234	0.07	-	MIGHT	N/A	8	0.02

Table 3 shows that only *shall* and *may* have positive keyness and have relatively higher percentage of occurrence in the two study corpora, while other modal verbs all showed negative keyness.

Following are some authentic examples of modal verbs' concordances within the private and public corpora.

(1) a. A lifeline **shall** also be fitted around the inside of the raft. (Public, SOLAS1.txt)
b. Reasonable notice of any such action **shall** be given to the defendant. (Private, CLC.txt)

(2) a. It should be noted that ships **may** have a need for reception of certain maritime safety information while in port. (Public, SOLAS2.txt)
b. Spaces which are linked by ducts of large cross-sectional area **may** be considered to be common. (Private, IBC.txt)

(3) a. The following ranges of visibilities of the light **might** be expected in given atmospheric conditions. (Public, SOLAS1.txt)
b. In the case of a heated cargo, carriage conditions **might** need to be established … (Private, IBC.txt)

(4) a. Administration **should** verify that ships are able to implement the

hours of rest … (Public, ISPS2.txt)

b. If security is furnished in several forms, these **should** be enumerated. (Private, CLC.txt)

(5) a. He **must** enter in the logbook the reason for failing to proceed to the assistance of the persons in distress. (Public, SOLAS1.txt)

b. … incidents involving losses of such substances from ships **must** be reported by the master or other person having charge of the ship concerned. (Private, IMDC.txt)

(6) a. The SSP should detail the security measures which **could** be taken by the ship … (Public, ISPS2.txt)

b. Any other loads not specifically addressed, which **could** have an effect on the cargo containment system, shall be taken into account. (Private, IGC.txt)

(7) a. Liquefaction **can** result in cargo shift. (Public, ISMBC2.txt)

b. … that the ship **can** survive the assumed flooding conditions. (Private, IGC.txt)

(8) a. … she **would** be able to see only the stern light of that vessel but neither of her sidelights … (Public, COLREG1.txt)

b. He may further avail himself of the defences (other than the bankruptcy or winding up of the owner) which the owner himself **would** have been entitled to invoke. (Private, CLC.txt)

(9) a. This resolution was adopted on 1 January 2010 and the amendments **will** enter into force on 1 July 2010. (Public, ISBM 1.txt)

b. This temperature **will** not result in unacceptable hull stresses. (Private, IGC.txt)

As seen in Table 3 and the above authentic examples (1)-(9), *might* appears in the key words list of the public law corpus but it does not appear in the key words list for private law. Therefore, the negative keyness of *might* may be used as another complementary distribution feature to distinguish two genres.

3.2 Key clusters

3.2.1 Four-word key clusters in complementary distribution

A cluster is a group of words which follows each other in a text (Scott 2015). Another common term for such repeated sequences of words is “recurrent combinations” (Altenberg 1998), “lexical bundles” (Biber, Johansson, Leech, Conrad, and Finegan 1999), “chains” (Stubbs and Barth 2003), or “n-grams”. “A key cluster, like a wordlist cluster, represents two or more words which are found repeatedly near each other. However, a key cluster only uses key words” (Scott 2015). In this study, we concentrated on four-word clusters which are flexible enough to occur across a number of different texts in the maritime institutional corpora; yet at the same time their frequencies are sufficiently manageable to allow for a detailed analysis. Biber *et al.* (1999) adopted lexical bundles only if they appear at least 10 times per million words. Considering the relative small size of our study corpora, we set the normalized cut-off points at five appearances. The number of key clusters that WordSmith generates depends on the choice of a significance value. The keyness of key clusters was calculated with a *p*-value of 0.000001 using log-likelihood. The smaller the *p*-value, the fewer clusters are found, i.e., the clusters are statistically more significant. This study used four-word clusters because four-word clusters provide not so many, but adequate numbers for this study, compared to the numbers of other types of clusters. For example, the private corpus produced 3,490 three-word clusters, 2,199 four-word clusters and 1,425 five-word clusters, whereas the public corpus produced 30,137 three-word clusters, 25,565 four-word clusters and 20,531 five-word clusters.

Table 4 (below) displays the top 10 four-word key clusters extracted from maritime institutional texts.

Table 4. Top 10 four-word key clusters of public/private law corpora

N	Public Law	Keyness	N	Private Law	Keyness
1	IN ACCORDANCE WITH THE	554.22	1	IN ACCORDANCE WITH ARTICLE	336.06
2	THE DATE ON WHICH	473.46	2	THE SECRETARY-GENERAL OF THE	280.04
3	WITH THE PROVISIONS OF	464.93	3	THE DATE ON WHICH	278.43
4	WITH THE REQUIREMENTS OF	455.14	4	SECRETARY-GENERAL OF THE ORGANIZATION	257.63

5	THE PROVISIONS OF THE	447.80	5	OR OTHER FINANCIAL SECURITY	235.23
6	OF THE PRESENT CONVENTION	415.98	6	INSURANCE OR OTHER FINANCIAL	235.23
7	FOR THE PURPOSE OF	383.23	7	RATIFICATION ACCEPTANCE APPROVAL OR	212.82
8	INTERNATIONAL CONVENTION FOR THE	364.59	8	ACCEPTANCE APPROVAL OR ACCESSION	212.82
9	IN ACCORDANCE WITH REGULATION	354.80	9	BY THE FOLLOWING TEXT	207.22
10	ENTRY INTO FORCE OF	332.78	10	THE PROVISIONS OF THIS	207.22
...	...	...	...	...	...
2249	SIDES OF THE SHIP	24.47	1434	THE OPERATION OF THE	24.26
2250	AS PART OF THE	-30.12			

With the given *p*-value, Wordsmith found 2,249 positive and one negative key cluster, *AS PART OF THE,* in the public law corpus, with a positive keyness ranging from 24.47 to 554.22 and the negative keyness, -30.12. In the private law corpus, only 1,434 positive key clusters were found, with a positive keyness ranging from 24.26 to 336.06. There is no occurrence of any negative key clusters. As a result, we may conclude that in maritime legal English, compared with general English, many clusters are overused, but very few are underused.

Among the top 10 key clusters, the three key clusters --- *WITH THE REQUIREMENTS OF* (455.14), *INTERNATIONAL CONVENTION FOR THE* (364.59), and *IN ACCORDANCE WITH REGULATION* (354.80) --- appear only in the public law corpus, whereas *OR OTHER FINANCIAL SECURITY* (235.23), *INSURANCE OR OTHER FINANCIAL* (235.23), and *BY THE FOLLOWING TEXT* (207.22) appear only in the private law corpus. Hence, these 4-word clusters can be used to distinguish public law from private law.

Moreover, we observed that in the top 10 key clusters, there are six prepositional phrases in the public law corpus, whereas only two prepositional phrases are in the private law corpus. If syntactic category preference affects different discourse roles, further discussion is needed. We will elaborate on this further in the following

section.

3.2.2 Relationship between structure and function classifications

Biber *et al.* (1999) propose structural classification of lexical bundles, i.e., NP-based category, PP-based category, and VP-based category, and describe their discourse roles across different genres. In this study, based on Biber *et al.* (1999), we classified the structures of significant four-word key clusters in the two study corpora. However, little consensus exists regarding the determination of the appropriate cut-off point. Biber (2006: 134) set the cut-off point at 0.004%, even though he pointed out that any of the bundles in his study occurred more than 0.02%. In the present study, we take a conservative approach, setting a relatively high frequency cut-off of 300 times in 1 million words (or 0.03%) so as to reduce the data to manageable quantities. The coverage of the corpus must also be considered; therefore, the chosen lexical bundles should appear in at least 20% of the study corpus. Biber and Barbieri (2007) mention that by using a normalized rate of occurrence, the bundles across sub-corpora of different sizes can be compared: to be considered a lexical bundle, a four-word sequence must recur at this rate, regardless of the size of the sub-corpus being analyzed. The result is listed in Table 5.

Table 5. Structural classification of significant four-word key clusters

Category	Public Law		Private Law		Examples
	Type	%	Type	%	
NP	21	52	48	48	*THE DATE ON WHICH*
PP	17	43	39	39	*IN ACCORDANCE WITH THE*
VP	2	5	13	13	*SHALL ENTER INTO FORCE*
	40	100	100	100	

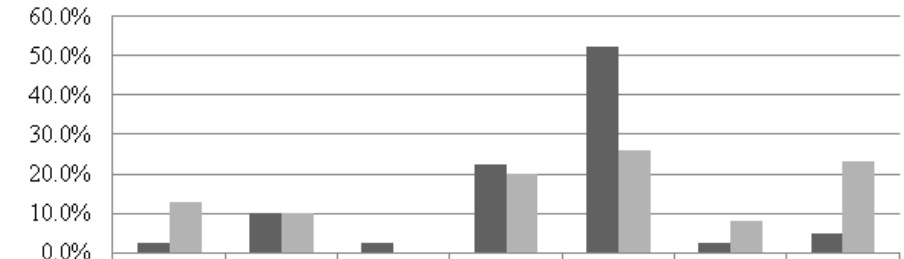
A. Stance Expressions denoting obligation and regularities, e.g., *shall be deemed to*.
B. Referential Expressions:
B1 Time Reference referring to specific times, e.g., *after the date of*.
B2 Quantity Specification specifying denoting quantities or amounts, e.g., *percent of the*.
B3 Specification of Intangible Framing Attributes that are often used to establish logical relationship in a text, e.g., *in accordance with the*.

C. Content Phrases:
C1 Legal Documents, e.g., *convention on civil liability*.
C2 Agents (people/institution), e.g., *insurer or other person*.
C3 Abstract Concepts, e.g., *the international monetary fund*.

From Table 5, we can see that in maritime institutional texts, NP-based and PP-based four-word key clusters take up a large proportion of the key clusters. In addition, the private law corpus contains more VP-based key clusters than the public law corpus.

Breeze (2013: 235) mentions that because of the specialized nature of legal documents, a large number of subject-specific noun phrases and prepositional phrases had been found, which referred to documents, institutions, people, procedures and theoretical concepts, termed "content phrases" (Pecorari 2009). Our analysis of lexical bundles is therefore based on Biber and Conrad's (1999) classification of classroom teaching and Breeze's (2013) classification of legal genres.

In our corpus, the significant four-word key clusters can be classified into three types of function classification, as shown in Figure 1.



	A	B1	B2	B3	C1	C2	C3
Public	2.5%	10.0%	2.5%	22.5%	52.5%	2.5%	5.0%
Private	13.0%	10.0%	0.0%	20.0%	26.0%	8.0%	23.0%

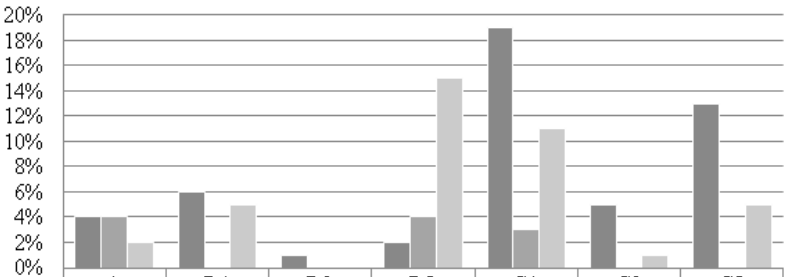
Figure 1. Function classification of significant four-word key clusters

We can see that the content phrases frequently used in legal documents take up a large part of discourse function due to the specialized features of the study corpora. In addition, private law incorporates a large proportion of the type A function since many *shall* related clusters are used. Moreover, a large amount of public law key clusters fall into the categories of B3 and C1 while the private law

corpus has more occurrences in types A and C3 category than the public law corpus.

Several former studies on lexical bundles have agreed with Biber *et al.*'s (1999) observation that instead of representing complete structural units, bundles tend to consist of syntactic fragments that extend across structural units, and their functional and structural distributions can go across different academic genres (Biber and Cortes 2004; Biber, Conrad, and Cortes 2004; Cortes 2004; Chen and Baker 2010; Simpson-Vlach and Ellis 2010).

Figure 2 shows the relationship between the structural and function classifications in the study corpora.



	A	B1	B2	B3	C1	C2	C3
NP	4%	6%	1%	2%	19%	5%	13%
VP	4%	0%	0%	4%	3%	0%	0%
PP	2%	5%	0%	15%	11%	1%	5%

Figure 2. Relationship between structural and function classification

From Figure 2, we can see that NP-based lexical bundles occur more frequently as content phrases, while the PP-based category occurs more as referential expressions. This is understandable since most of the referential expressions consist of prepositional phrases.

3.3 Key semantic domains in complementary distribution

Most linguistic analyses of frequently-used words have been based either on parts of speech (word class) or on syntactic categories. With the application of more advanced searching techniques such as the Wmatrix system utilized in the present study, semantic domains can be accounted. Wmatrix, designed by Rayson (2008), is

an automatic tagging software that

assigns a group of key words a semantic field (domain) tag, extracting key domains by applying the keyness calculation to tag frequency lists. According to Rayson (2008: 519), the combination of the key words and key semantic domains is shown to allow macroscopic analysis (the study of the characteristics of whole texts or varieties of language) to inform the microscopic level (focusing on the use of a particular linguistic feature).

Tables 6 and 7 show the top 15 key semantic domains of this study's public and private law corpora sorted by the value of Log-likelihood while comparing with the reference BNC-written-informative corpus. O1 and O2 are observed frequency in the study corpus and in the reference corpus respectively. The values listed under %1 and %2 show relative frequencies in the texts. Examples of the assigned semantic field and its correspondent semantic category are also shown in the tables.

Table 6. Top 15 key semantic domains of the public law corpus

N	Items	Semantic Fields	Log-likelihood	O1	%1	O2	%2	Examples
1	M4	Sailing, swimming, etc.	12,622.41	6,898	2.18	678	0.09	ship(s), vessel(s), lifeboat(s), marine, crew, navigation
2	T1.1.3	Time: Future	5,422.6	6,538	2.07	3,435	0.46	future, imminent
3	N1	Numbers	4,232.29	13,350	4.22	14,171	1.9	1, 1997, iii, three, 4.1, numeral
4	G2.1	Law and order	3,644.49	4,749	1.46	2,770	0.36	regulation(s), law, privilege
5	O2	Objects generally	3,428.07	6,457	2.04	5,055	0.68	cargo, machinery, log, grinder
6	Q1.2	Paper documents and writing	2,008.12	3,840	1.21	3,039	0.41	certificate, list, signatory, receipt, document(s)
7	S7.4+	Allowed	1,834.38	2,100	0.66	1,040	0.14	approved, ratification, permit
8	M6	Location and direction	1,779.22	6,374	2.01	7,174	0.96	position, end, transfer, direction, south, sideways,
9	N3.6	Measurement: Area	1,751.93	923	0.29	76	0.01	gram, square_yard, space(s)
10	O1.2	Substances and materials:	1,609.93	1,749	0.55	813	0.11	oil(s), water(s), liquid(s),

		Liquid						waterline, fluid, humidity
11	A1.2+	Suitable	1,375.24	1,251	0.4	454	0.06	appropriate, relevant, suitable
12	A15+	Safe	1,288.49	774	0.24	110	0.01	guard, safety, guardrails, safe
13	N3.5	Measurement: weight	1,241.93	1,011	0.32	304	0.04	tonnage, weight, kg, lbs, mg
14	N5.1+	Entire; maximum	1,025.47	4,059	1.28	4,766	0.64	in total, all, each, gross, full, filling, full-scale, utmost
15	A6.2+	Comparing: Usual	950.27	2,271	0.72	2,078	0.28	standard, normally, basic, regular, routine, naturally

Table 7. Top 15 key semantic domains of the private law corpus

N	Items	Semantic Fields	Log-likelihood	O1	%1	O2	%2	Examples
1	G2.1	Law and order	3,250.15	1,498	2.96	2,700	0.35	regulation(s), law, article(s)
2	N1	Numbers	1,767.21	2,629	5.18	14,171	1.9	quarter, 1, 4.1, iv
3	A6.2+	Comparing: Usual	1,215.52	779	1.54	2,078	0.28	standard, normally, basic, regular, routine,
4	T1.1.3	Time: Future	1,102.57	951	1.87	3,435	0.46	future, be about to
5	M4	Sailing, swimming, etc.	1,090.44	461	0.91	678	0.09	ship(s), vessel(s), lifeboat(s), marine, crew
6	Q3	Language, speech and grammar	993.58	562	1.11	1,286	0.17	language(s), word(s), expression(s), sentence(s)
7	S7.4+	Allowed	855.52	471	0.93	1,040	0.14	approved, ratification, permit
8	A1.1.2	Damaging and destroying	692.23	336	0.66	620	0.08	damage, wreck, harmful, victim(s), accident, collision
9	Q1.2	Paper documents and writing	649.21	700	1.38	3,039	0.41	certificate, list, signatory, receipt, document
10	S7.1-	No power	637.35	316	0.62	602	0.08	servants, surrender, depend
11	G1.1	Government	628.89	728	1.43	3,330	0.45	authority, civil, government
12	A2.2	Cause and Effect / Connection	400.55	667	1.31	3,836	0.51	effect(s), caused, result(s), reasons, consequence(s)
13	N5.1+	Entire; maximum	366.93	747	1.47	4,766	0.64	all, each, every, gross
14	W5	Green issues	308.01	133	0.26	203	0.03	pollution, environment, nature
15	M6	Location and direction	291.27	931	1.83	7,174	0.96	position, end, transfer, direction, south, sideways,

Through comparison of Table 6 and Table 7, nine overlapped key semantic domains are found across the two study corpora. These nine domains are the typical semantic fields of maritime institutional texts. Numbers (N1) includes number terms (e.g., *cardinal, ordinal, faction*, etc.). Location and direction (M6) is a label depicting position of / point of reference for X, e.g., *ashore, backward, adjacent to*,

etc. Time: General: Future (T1.1.3) includes words such as *future, imminent*, etc. The modal words *shall* and *will* are also counted inside this category. Entire; maximum (N5.1+) is a label depicting maximal / maximum quantities which are used to modify the objects generally. Since these conventions are closely associated with the maritime field, the occurrence of sailing, swimming, etc. (M4) is not unexpected. Law and order (G2.1) includes terms relating to legal systems. Comparing: Usual (A6.2+) includes comparative terms a denoting level of normality, including *standard, basic, regular*, etc. Legal conventions may also include words in key semantic domains of Paper documents and writings (Q1.2) and Allowed (S7.4+). These nine domains are shared by both private and public law corpora.

As for the differences, the public law corpus contains the conventions for the registration of vessels, safety of ships and safety of navigation, control of shipping operations and so on. Thus, words in Measurement (N3.6 and N3.5), Safe (A15+) and Suitable (A1.2+) are used. International private maritime law is concerned with legal relationships between individuals or groups of individuals such as co-operatives, companies, etc. Its primary purpose is the protection of individuals' interests such as the acquisition and transfer of the ownership of vessels, charter-parties, and bills of lading. The conventions included in this corpus are often related to liability for damage and its compensation, therefore words related to Government (G1.1), Damaging and destroying (A1.1.2) and Cause and Effect / Connection (A2.2) are used. A key semantic domain of Green issues (W5) is also high ranked in the private corpus, due to the fact that accidents at sea may lead to oil leaking, pollution to the environment.

Interestingly enough, key semantic fields of Measurement: Area (N3.6) and Safe (A15+) are in complementary distribution since they appear only in the list of key semantic domains of the public law corpus but never appear in the list of key semantic domains in the private law corpus. These two semantic fields can also be a way to distinguish public law from private law.

4. Conclusion

This paper discusses how typical linguistic features of maritime institutional texts were identified by comparing a self-built maritime legal English corpus with a

general English corpus. We have focused particularly on the differences in English between written international public and international private maritime laws with regard to lexico-grammatical and semantic and discourse features such as key words, key clusters and key semantic domains.

Several interesting and important findings were observed. First, in the analysis of key words, we found that *regulation / administration vs. liability / compensation* are in complementary distribution in the public law corpus and the private law corpus respectively. Among modal auxiliary verbs, only *shall* and *may* have positive keyness and have higher frequency in the two corpora, while other modal verbs all show negative keyness. Second, in the analysis of key clusters, complementary four-word clusters can be used to distinguish public law from private law. Although having very high keyness (455.14; 364.59; 354.80) respectively, *WITH THE REQUIREMENTS OF*, *INTERNATIONAL CONVENTION FOR THE*, and *IN ACCORDANCE WITH REGULATION* are the four-word clusters appearing only in the public law corpus, whereas *OR OTHER FINANCIAL SECURITY*, *INSURANCE OR OTHER FINANCIAL*, and *BY THE FOLLOWING TEXT* are the four-word clusters that are specifically used in the private law corpus with the keyness of 235.23, 235.23 and 207.22 respectively. These four-word clusters can be used to distinguish public law from private law. Moreover, the further analysis of the significant four-word key clusters also shows the differences between structure and function classifications in the two study corpora. NP-based lexical bundles occur more frequently as content phrases, whereas PP-based category more frequently as referential expression.

Finally, Measurement: Area (N3.6) and Safe (A15+) appear only in the list of key semantic domains of the public law corpus, never in the similar list for the private law corpus. Hence, key semantic fields of Measurement: Area and Safe are in complementary distribution in the two corpora: these two semantic fields distinguish public law from private law.

These linguistic features, discussed from the view of keyness, may provide non-legal practitioners and non-specialist readers to discover and describe underlying parameters that best distinguish between legal registers or genres. We further believe that the methodology for ascribing differences between the public and the private law, as proposed in this paper, may be effective for general legal documents as well as other maritime institutional texts.

References

Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations. In Anthony Paul Cowie (ed.), *Phraseology, theory, analysis, and applications*, 101-122. Oxford, UK: Oxford University Press.

Asensio, Roberto Mayoral. 2003. *Translating official documents*. Manchester, UK: St. Jerome.

Bhatia, Vijay K. 1993. *Analyzing genre: Language use in professional settings*. London, UK : Longman.

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

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. 1999. Lexical bundles in conversations and academic prose. In Hilde Hasselgård and Signe Oksefjell (eds.), *Out of corpora: Studies in honour of Stig Johansson*, 181-190. Atlanta, GA: Rodopi.

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*. London, UK: Longman.

Bocanegra-Valle, Ana. 2013. Maritime English. In Carol. A. Chapelle (ed.), *The encyclopedia of applied linguistics*, 3570-3583. Oxford: Wiley-Blackwell.

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 and Technology* 14(2): 30-49.

Conrad, Susan and Douglas Biber. 2004. The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica* 20: 56-71.

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

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

Goh, Gwang-Yoon. 2011. Choosing a reference corpus for keyword calculation. *Linguistic Research* 28(1): 239-256.

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

Hafner, Christoph and Christopher N. Candlin. 2007. Corpus tools as an affordance to learning in professional legal education. *Journal of English for Academic Purposes* 6(4):

303-318.

Hong, Shin-Chul and Se-Eun Jhang. 2010. The compilation of a Maritime English corpus for ESP learners. *Korean Journal of English Language and Linguistics* 10: 963-985.

Iber, Jean-Jacques. 2001. A concordance- and genre- informed approach to ESP essay writing. *ELT Journal* 55: 14-20.

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

Johns, Ann. M. 2013. The history of English for specific purposes research. In Brian Paltridge and Sue Starfield (eds.), *The handbook of English for specific purposes*, 5-30. Chichester, UK: John Wiley and Sons, Ltd.

Lee, Sung-Min. 2016. *Network analysis of Maritime English corpus with multi-word compounds: Keyword networks and collocation networks*. Unpublished Ph.D. Thesis, Korea Maritime and Ocean University.

Maley, Yon. 1994. The language of the law. In John Gibbons (ed.), *Language and the law*, 3-50. Harlow, UK: Longman.

Northcott, Jill. 2013. Legal English. In Brian Paltridge and Sue Starfield (eds.), *The handbook of English for specific purposes*, 213-226. Chichester, UK: John Wiley and Sons, Ltd.

Orts-Llopis, M. Angeles. 2009. Legal genres in English and Spanish: Some attempts of analysis. *Ibérica* 18: 119-130.

Orts-Llopis, M. Angeles. 2014. Contractual commitment, or obligation? The linguistic interactions in Charter Parties. In Ruth Breeze, Maurizio. Gotti, and Carmen. Sancho-Guinda (eds.), *Interpersonality in legal genres series: linguistic insights*, 87-112. Frankfurt am Main, Germany: Peter Lang.

Pecorari, Diane. 2009. Formulaic language in biology. A topic-specific investigation. In Susan Hunston, Maggie Charles, and Diane Pecorari (eds.), *Academic writing: At the interface of corpus and discourse*, 91-105. London, UK: Continuum.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London, UK: Longman.

Rayson, Paul. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics* 13(4): 519-549.

Rayson, Paul. 2009. Wmatrix: A web-based corpus processing environment. Computing Department, Lancaster University [EB/OL]. [http://ucrel.lancs.ac.uk/wmatrix3.html] (last accessed March 2016).

Rayson, Paul, Dawn Archer, Scott Piao and Tony McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic Labeling for NLP Tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 25th May, Lisbon, Portugal, 7-12. Paris, France: European Language Resources Association.

Šarčević, Susan. 1997. *New approach to legal translation*. Hague, Netherlands: Kluwer Law International.

Scott, Mike. 1997. PC analysis of key words and key key words. *System* 25(2): 233-245.

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

Scott, Mike. 2015. *Wordsmith tools help. Liverpool: Lexical analysis software*. Retrieved from [http://www.lexically.net/downloads/version6/HTML/index.html?getting_started.html].

Simpson-Vlach, Rita and Nick. C. Ellis. 2010. An academic formulas list (AFL). A*pplied Linguistics* 31(4): 487-512.

Sinclair, John. 1996. The search for units of meaning. *Textus* 9(1): 75-106.

Stubbs, Michael and Isabel Barth. 2003. Using recurrent phrases as text-type discriminators. A quantitative method and some findings. *Functions of Language* 10(1): 61-104.

Williams, Raymond. 1976/1983. *Keywords: A vocabulary of culture and society*. London, UK: Fontana.

Appendix

Retrieved September 10, 2015 from [http://www.imo.org/en/About/Conventions/List Of Conventions/Pages/Default.aspx]

Public maritime institutional corpus

A. Conventions relating to maritime safety/security and ship/port interface

1. International Convention for the Safety of Life at Sea (SOLAS), 1974 Protocol of 1978 and the Protocol of 1988 and the Amendments to the Annex (1995).
2. International Convention on Standards of Training, Certification and Watchkeeping for Seafarers (STCW), 1978
3. Convention on the International Regulations for Preventing Collisions at Sea (COLREG), 1972
4. International Convention on Tonnage Measurement of Ships (TONNAGE), 1969
5. Convention on Facilitation of International Maritime Traffic (FAL), 1965
6. International Convention on Load Lines (LL), 1966
7. International Convention on Maritime Search and Rescue (SAR), 1979
8. International Convention for Safe Containers (CSC), 1972
9. Special Trade Passenger Ships Agreement (STP), 1971 and Protocol on Space Requirements for Special Trade Passenger Ships, 1973
10. International Convention on Standards of Training, Certification and Watchkeeping for Fishing Vessel Personnel (STCW-F), 1995
11. Convention on the International Maritime Satellite Organization (IMSO C), 1976
12. Convention for the Suppression of Unlawful Acts Against the Safety of Maritime Navigation (SUA), 1988
13. Protocol for the Suppression of Unlawful Acts Against the Safety of Fixed Platforms located on the Continental Shelf, 2005. The Torremolinos International Convention for the Safety of Fishing Vessels (SFV), 1993

B. Conventions relating to prevention of marine pollution

1. International Convention Relating to Intervention on the High Seas in Cases of Oil Pollution Casualties (INTERVENTION), 1969
2. Convention on the Prevention of Marine Pollution by Dumping of Wastes and Other Matter (LC), 1972
3. International Convention on Oil Pollution Preparedness, Response and Co-operation (OPRC), 1990
4. Protocol on Preparedness, Response and Co-operation to pollution Incidents by Hazardous and Noxious Substances, 2000 (OPRC-HNS Protocol)

5. International Convention on the Control of Harmful Anti-fouling Systems on Ships (AFS), 2001
6. International Convention for the Prevention of Pollution from Ships, 1973, as modified by the Protocol of 1978 relating thereto and by the Protocol of 1997(MARPOL)
7. International Convention for the Control and Management of Ships' Ballast Water and Sediments (BWM), 2004
8. The Hong Kong International Convention for the Safe and Environmentally Sound Recycling of Ships, 2009

Private maritime institutional corpus

Conventions covering liability and compensation

1. International Convention on Civil Liability for Oil Pollution Damage (CLC), 1969
2. 1992 Protocol to the International Convention on the Establishment of an International Fund for Compensation or Oil Pollution Damage (FUND 1992)
3. Convention relating to Civil Liability in the Field of Maritime Carriage of Nuclear Material (NUCLEAR), 1971
4. Athens Convention relating to the Carriage of Passengers and their Luggage by Sea (PAL), 1974
5. Convention on Limitation of Liability for Maritime Claims (LLMC), 1976
6. International Convention on Liability and Compensation for Damage in Connection with the Carriage of Hazardous and Noxious Substances by Sea (HNS), 1996
7. International Convention on Civil Liability for Bunker Oil Pollution Damage, 2001
8. International Convention on Salvage (SALVAGE), 1989
9. Nairobi International Convention on the Removal of Wrecks, 2007

Wenyu Lu
Dalian Maritime University[a] · Korea Maritime and Ocean University[b]
Department of English[a]
1 Linghai Rd, Ganjingzi Qu, Dalian Shi, Liaoning Sheng 116000, China
Department of English Language and Literature[b]
727 Taejong-ro, Yeongdo-gu, Busan 49112, Korea
E-mail: wenyu-lu@dlmu.edu.cn

Sung-Min Lee
Korea Maritime and Ocean University
Department of English Language and Literature
727 Taejong-ro, Yeongdo-gu, Busan 49112, Korea
E-mail: roy7942@hanmail.net

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

Received: 2016. 10. 17.
Revised: 2017. 03. 02.
Accepted: 2017. 03. 02.