Visualization of Collocational Networks: Maritime English Keywords

Se-Eun Jhang and Sung-Min Lee
(Korea Maritime and Ocean University)

Jhang, Se-Eun and Lee, Sung-Min. (2013). Visualization of Collocational Networks: Maritime English Keywords. *Language Research* 49.3, 781-802.

The purpose of this paper is to explore and then visualize collocational networks of the most frequent maritime English synonyms through network analysis. To achieve this goal, we used WordSmith Tools to extract keywords from the Maritime English Corpus (MECO) Ⅱ and then we compared them with a general English corpus, the British National Corpus Baby. We discuss two pairs of near-synonyms: *maritime-marine* and *ship-vessel* from among very highly frequent keywords in the MECO Ⅱ. We used Mutual Information 3 to discover significant collocations in order to find collocational patterns. The meanings of collocates of near-synonyms were grouped in several semantic fields. In order to explore language networks of maritime vocabularies, we used the social network analysis tools, NetMiner and UCINET. We found that collocates of near-synonyms were quite different. After we extended our work to all collocations, we found that the entire network of all collocations also shows distributions and characteristics similar to those of our target networks.

Keywords: Maritime English, keywords, near-synonyms, collocation, network analysis, visualization, Mutual Information 3, WordSmith Tools, NetMiner, UCINET

1. Introduction[1)]

Network analysis has been used to describe a large number of sys-

1) Earlier versions of this paper were presented at the New Korean Association of English Language and Literature (May, 2013), the International Conference on English Linguistics (July, 2013), and the Korean Association of Language Sciences (August, 2013). These earlier versions have been revised as regards data and methodology. We would like to express our special thanks to the audience at the conferences, Dr. Mike Scott, and to three reviewers for their valuable comments. Any remaining errors, however, are our own responsibility.

tems in the real world, including the World Wide Web, biomedical studies, and human organizations. It powerfully contributes to evaluating relationships among abstract elements, people, and knowledge (Watts and Strogatz 1998, Scott 2000, Newman 2001, Barabasi 2002, Christakis and Fowler 2010, Lee 2012). Network studies have asked how network structures are constructed and what the structures really mean.

Similarly, in the field of linguistics, researchers have studied the meaning of words in accordance with their co-occurrence relationship in network concepts, namely, "collocation", because the meaning of a word can be better explained in the context of interwoven word groups. There is general agreement that individual words and their co-occurrences contribute to shaping the meaning of words (Sinclair 1991, Lewis 1997, Schmitt 2000, Nation 2001). Notably, Sinclair, Jones, and Daley (2004: 10) defined collocation as "the co-occurrence of two items within a specified environment". Sinclair and his colleagues discovered the notion of statistical collocation through exhaustive empirical testings and studied how to find better window-spans or positions of co-occurrences.

Thus, recent corpus linguistics studies have incorporated the notion of collocations into their visualization. McEnery (2006) showed networks of keywords linked through common collocates in a general English corpus for an explanatory account of swearing. Using the British National Corpus (BNC) compiled in the early 1990s, Beavan (2008) presented "collocate clouds" with the collocates of lexical items by ordering them alphabetically and by altering their font size and brightness. As for studies using specialized corpora, Williams (1998) explored collocational networks looking at patterns of biology terms in a corpus of plant biology articles. Stuart and Botella (2009) studied knowledge networks of specific science discourse communities. These studies demonstrate that collocational networks are a useful means of exploring complex relationships between lexical items.

Considering previous studies,[2] we believe that the collocational networks approach can enable us to analyse large and complex data.

2) The notion of collocation means not only statistically significant co-occurrence with node words (Sinclair et al. 2004), but also psychological reality (Hoey 1991). Hoey described the special textual patterns as a network of links with his term "priming", meaning explicit memory effect.

Visualized data will tell us whether the near-synonymous words share similar or different meanings. To apply this methodology, we used Maritime English as our target corpus as it is used in maritime discourse communities because such a specialized corpus has received little attention. Maritime English is defined as an official language within the international maritime community, contributing to the safety of navigation and the facilitation of seaborne trade (Trenker 2000, IMO 2009, Bocanegra-Valle 2012).

We will try to provide answers to the following four research questions raised in this study. (1) How do we extract keywords from the Maritime English Corpus built for English for Specific Purposes (ESP)? (2) How do we find significant collocations from our target vocabularies of *maritime-marine* and *ship-vessel?* (3) What do nodes[3)] and their co-occurring node to the left and right really when corpus linguistic data are used in social network analysis? (4) How do we visualize collocational networks of *maritime-marine* and *ship-vessel?*

This paper is organized as follows: Section 2 addresses previous research related to the topic. Section 3 discusses data and methodology for analysis. Section 4 illustrates collocational network structures visualized from our corpus data. Section 5 summarizes our findings.

2. Literature Review for Language Network and Its Visualization

Recently, human language has been studied within the framework of complex network analysis. Masucci and Rodgers (2006) studied network properties of Orwell's *1984.* They treated the whole text as a network where each word is a node and two words are linked when they co-occur, as seen in Figure 1.

3) We used the terms "network", "node", and "link" in accordance with computer science terminology. These terms are called "graph", "vertex", and "edge" in mathematics.

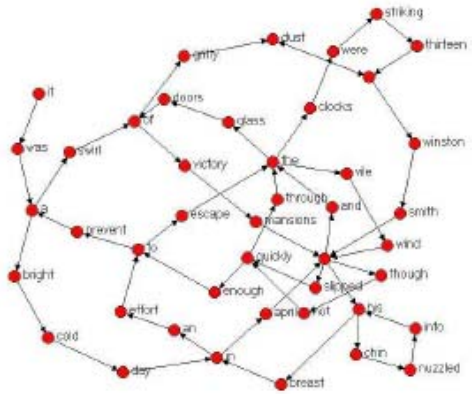

Figure 1. Language network for the first 60 words of Orwell's 1984.

Masucci and Rodgers analysed the properties of the nearest neighbors and clustering coefficients and found that they show the characteristics of power law, also known variously as Zipf's law or the Pareto distribution. There are a few similar studies: Zhou et al. (2008) studied Chinese language networks from "The People's Daily" corpus using complex network theory. They built two different networks based on different criteria to define link relations. Liang et al. (2009) studied collections of Chinese and English essays, novels, popular science articles, and news reports. They found diameter, average degree, degree distribution, clustering coefficients, and average shortest path length in the Chinese and English languages.

In corpus linguistics, there are a number of researchers who analysed corpus data relying on visualization techniques (McEnery 2006, Beavan 2008, Rayson and Mariani 2009, Stuart and Botella 2009, Lee and Lee 2010, Jung and Kang 2011, Scott 2012).[4] McEnery (2006) employed a visualization technique drawing collocational networks, showing keywords linked by common collocates, as seen in Figure 2.

4) For Korean research, see Lee and Lee (2010) for language networks in the Yonsei Korean Dictionary, and Jung and Kang (2011) for co-occurrence networks of family nouns in the newspapers.

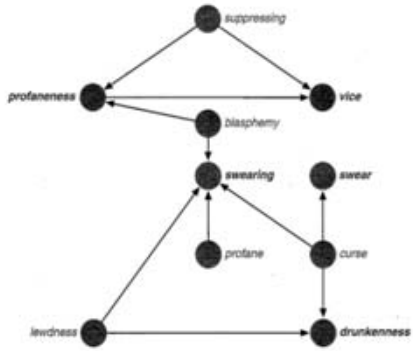

Figure 2. Collocational networks visualization.

Regarding specialized corpus studies, William (1998) explored specific corpora containing biology articles in order to demonstrate collocational networks and to find meanings of head words surrounded by collocations. Stuart and Botella (2009) analyzed keywords and clusters in terms of their distributions across text plots and discipline levels. Their results indicate that researchers in science discourse university communities share some keywords and clusters, as seen in Figure 3.

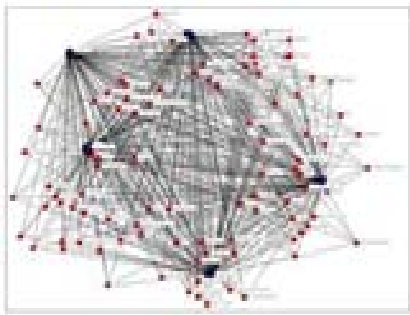

Figure 3. A network example of keywords per document section.

Recently, using WordSmith Tools, Scott (2012) created word clouds[5] in which frequent words are printed larger and take center position in order to point out frequent words, as seen in Figure 4.

5) In similar research, Rayson and Mariani (2009) extracted keywords from corpus linguistics conference articles and created keyword clouds in which words with higher keyness are printed larger. This work leads researchers to visually identify the gradual change in trends for corpus studies.

Figure 4. Concordance and frequent words cloud visualization.

3. Data and Methodology

3.1. Statistical Data

This paper deals with Maritime English, studying collocates of two pairs of near-synonyms appearing as four highly frequent keywords in the Maritime English Corpus (MECO) II.[6] The MECO II is composed of over 1.4 million words including academic prose, periodicals, documents, and communications.[7] Its detailed statistical data are shown in Table 1.

As a reference corpus, we used the BNC Baby because the corpus has been regarded as representing general native English. The BNC Baby is a balanced corpus which was selected from the 100 million words-sized BNC. The result of careful sampling by experts, the BNC Baby consists of over 4 million words, including academic prose, newspapers, fiction, and conversations as its sub-corpora.

6) The MECO originally consisted of about 1 million words: journals, periodicals, documents, and communications compiled by Hong and Jhang (2010). The one million words-sized MECO was used for several studies about Maritime English (Hong and Jhang 2010, Jhang and Byun 2011, Jhang and Parent 2011, Jhang and Lee 2013a, 2013b).

7) 400,000 words from both 120,000 spoken and 280,000 written texts were added later. Spoken data consist of seaman's interviews, maritime lectures, and VTS communications. Written data consist of maritime-related documents.

Table 1. Basic statistical information of corpora

Statistics	MECO II	BNC Baby
Text file	98	182
Tokens (running words) in text	1,526,597	4,104,735
Tokens used for word list	1,402,389	4,056,526
Types (distinct words)	35,086	76,064
Standardised Type/Token ratio	35.9	41.5
Sub-corpora	Academic prose Periodicals Documents Communications	Academic prose Newspapers Fictions Conversations

3.2. Methodology for Keywords Extraction

The concept of a keyword[8] means "a word which occurs with unusual frequency in a given text" defined by Scott (1997). One of the most reliable statistical tests is a log-likelihood test, as proposed by Rayson (2003). He argued for the justification of log-likelihood and adopted it in his subsequent research (e.g., Rayson 2008). In addition, Scott (2010) proposed the concept of keywords using log-likelihood. In this sense we followed Dunning's (1993) log-likelihood test because it gives a better estimate of keyness, especially when contrasting long texts against the reference corpus (Scott 2012). Using the keywords tool WordSmith Version 6, we selected the MECO II as a target corpus and the BNC Baby as a reference corpus. Regarding statistical criteria, we set a relatively low p-value threshold, 0.000001 (1 in one million), to produce more reliable raw data and finally obtained 4,130 significant keywords. Among these outcomes the top 20 keywords are shown in Figure 5.

8) There are a number of statistical methods for finding keywords such as Yule's different coefficients, Pearson's chi-square, log-likelihood, normalised ratio, and Fisher's Exact tests.

KeyWo

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	Texts	RC. Freq.	RC. %	Keyness
1	SHIP	4,944	0.32	79	171		11,280.21
2	SHIPS	3,632	0.24	68	36		8,922.49
3	VESSEL	3,397	0.22	71	26		8,412.64
4	MARITIME	3,296	0.22	65	5		8,370.77
5	SHALL	4,050	0.27	49	976	0.02	6,061.78
6	CARGO	2,406	0.16	65	24		5,907.97
7	PORT	2,428	0.16	80	95		5,470.50
8	VESSELS	2,087	0.14	68	25		5,089.23
9	SHIPPING	1,946	0.13	73	21		4,764.50
10	IMO	1,668	0.11	44	0		4,271.16
11	INTERNATIONAL	2,539	0.17	72	463	0.01	4,222.23
12	CONVENTION	1,859	0.12	46	82		4,134.37
13	OF	54,378	3.56	97	100,766	2.54	3,988.16
14	OIL	1,914	0.13	55	252		3,507.66
15	SAFETY	1,836	0.12	67	254		3,320.39
16	THE	100,263	6.56	98	209,735	5.29	3,273.43
17	REGULATION	1,359	0.09	51	35		3,175.48
18	MARINE	1,459	0.10	64	102		3,048.53
19	TRANSPORT	1,456	0.10	51	186		2,688.99
20	CODE	1,287	0.08	60	98		2,651.14

Figure 5. Top 20 keywords calculated by log-likelihood.

Figure 5 shows the top 20 keywords in order of keyness values. There are function words such as *shall, of,* and *the* on this list. These words indicate certain characteristics of Maritime English. For example, Jhang and Byun (2011) and Jhang and Lee (2013b) pointed out that *shall* was found as keywords within the top 30, because a number of conventions, regulations, and codes used in maritime communities employ *shall*-related expressions (e.g., *shall be provided with, shall be deemed to,* etc.). The function words *of* and *the* are also highly frequent because Maritime English contains more noun phrases (e.g., *the gross tonnage of*, *the east coast of*, etc.) than general English. From the keyword list of 20 above, we chose *maritime-marine* and *ship-vessel* as near-synonyms.

3.3. Methodology for Statistical Tests to Obtain Collocations

We needed a statistical method to find collocations of the near-synonyms in question. In order to focus on finding statistical collocations of each of the individual keywords, we decided to use Mutual Information (MI) 3 as a statistical measure[9] based on information theory.

9) Another statistical measure, hypothesis testing of differences (HTOD), was proposed

Our reason is that MI or MI3 was devised to spot not only words that occur adjacently but also words that co-occur in a text (Walter 2010).[10] In order to obtain more reliable results, Church and Hanks (1990) and Church et al. (1991) proposed MI for finding associations between two words. If two events x and y are independent, then I (x, y) = 0. However, MI tends to give too much weight to infrequent words. MI3 was proposed to mitigate this problem. It is computed as follows:

$$I(x,y) = \log^2 \frac{P(x,y)^3}{P(x)P(y)}$$

MI3 is calculated by dividing observed frequencies of the co-occurring words by expected frequencies of the co-occurring words within specific spans, taking the logarithm to the base 2 of the outcome. By adding 'cubing' observed frequencies, MI3 made it possible to give more weight to high frequencies than to low frequencies (Oakes 1998: 171-172).

In order to find which test suits our study, we calculated collocations from MI and MI3. Table 2 shows how these methods produce different collocates in the case of *maritime* within the top 20 ranks.

Table 2. Comparison of MI and MI3 collocations of *maritime* within 20 ranks

	MI	Frequency	MI3	Frequency
1	FOR	296	ENGLISH	541
2	CAN	19	FOR	296
3	MUST	13	TRANSPORT	464
4	YEARS	11	THE	1,761
5	STATES	10	REVIEW	313
6	INSTITUTES	9	OF	1,227
7	MANUFACTURERS	8	INTERNATIONAL	412

in the earlier version of this paper. In the present paper, we substituted MI3 for HTOD because two reviewers pointed out that HTOD is suitable for comparison of collocates of near-synonyms (Manning and Schütze 1999), not for finding their collocates.

10) Dr. Mike Scott wrote his opinion about statistical measures at wordsmithtools@ googlegroups on November 11, 2013, saying that he has found MI3 useful himself. Numerous corpus studies have used MI or MI3 for finding collocations (Walter 2010).

	MI	Frequency	MI3	Frequency
8	NATIONS	27	AND	960
9	GOVERNMENTS	5	SAFETY	332
10	PILOTS	4	ORGANIZATION	216
11	PAPERS	4	SECURITY	226
12	CONTRACT	3	IN	559
13	OFFICERS	4	UNIVERSITY	119
14	ZEEVAARTSSCHOOL	2	COMMITTEE	121
15	COMPENSATION	2	TO	450
16	WARNEMUNDE	2	TEACHING	71
17	THEN	2	IMO	133
18	JUDITH	2	A	284
19	DAVIDS	1	EDUCATION	58
20	FOURNIE	1	NATIONS	27

From Table 2 it was evident that most of the collocates based on MI have fewer frequencies, even including one or two co-occurrences compared with the collocates from MI3. Moreover, 14 words from MI have less than 10 frequencies and the top frequency word came 296 times, whereas 17 words in MI3 have more than 100 frequencies and the top frequency word is 1,761. In addition, collocates from MI have fewer grammatical words than MI3 collocates. There are only three functional words from MI: *for, can*, and *must*, but there are seven functional words from MI3: *for, the, of, and, in, to*, and *a*. Even though MI and MI3 have been regarded as useful tools to identify significant collocations, the two tests showed different outcomes. Considering these results, we decided to choose MI3 as the better statistical test because the collocates from MI3 seem closer to the intuitive relation containing higher frequent content words in the case of maritime vocabulary.

Using the concord tool in the WordSmith Tools,[11] we investigated collocates within four spans to the left and the right regarded as significant collocations according to Sinclair et al. (2004). Collocates of each of four keywords were found: (a) *maritime* (1,721), (b) *marine*

11) After activating the Concord tool, we clicked the collocates option and then moved to compute menu at the top and clicked relationships where we were able to choose the statistical test method.

(994), (c) *ship* (2,335), and (d) *vessel* (1,712). Among these collocates we selected 50 high MI3 scored collocations.[12)]

3.3.1. *maritime* and *marine*

Collocates of *maritime* and *marine* were extracted by MI3 within four spans to left and right words and listed in alphabetical order. In Table 3 there are three common collocates: *industry, services,* and *transport* in the second row and 47 different collocates in the third row.

Table 3. Collocations of *maritime* and *marine* calculated by the MI3 test

(-4 maritime +4)		(-4 marine +4)	
INDUSTRY, SERVICES, TRANSPORT (3)			
ACADEMY	LANGUAGE	ACCIDENT	MAMMAL
ADMINISTRATION	LAW	ACCIDENTS	MANAGEMENT
ADMINISTRATIONS	LEARNING	ACT	MEMORIAL
ADOPTED	LECTURER	ARCHITECTS	MERCHANT
BUREAU	L'ORGANISATION	AVIATION	NEW
CENTRE	MOBILE	BIOSAFETY	OBSERVATION
COLLEGE	NATIONAL	CASUALTIES	PHRASES
COMMITTEE	NATIONS	CASUALTY	POLLUTANTS
COMMUNITY	ORGANIZATION	CENSUS	POLLUTION
CONSTANTA	RESCUE	CHANTS	PRACTICE
CONSULTATIVE	REVIEW	CLAIMS	PREVENTION
DISTRESS	SAFETY	CONTRACT	PROTECTION
EDUCATION	SEARCH	DUMPING	REFRIGERATION
ENGLISH	SECTOR	ECOSYSTEMS	RENEWABLE
GENERAL	SECURITY	ENERGY	RESOLVE
GENRES	STUDIES	ENGINEERING	SCIENCE
GLOBAL	TEACHERS	ENVIRONMENT	SHIPBUILDING
GOVERNMENTAL	TEACHING	ENVIRONMENTAL	SPILLS
IMO	TRAINING	EVACUATION	STANDARD
INSTITUTE	TRANSPORTATION	EVERGREEN	STRIKING
INSTITUTIONS	UNIVERSITY	INCIDENT	SURVEY
INTERNATIONAL	WORLD	INCIDENTS	SURVEYS
INTERNATIONALE	YEARS	INVASIONS	TECHNOLOGY
LABOUR	(47)	INVESTIGATION	(47)

We found that collocates of both *maritime* and *marine* deal with certain semantic fields. Common collocates are *industry, services,* and *transport*. These words are shipping industry terms. Regarding independent collocates, the first group is associated with accident-related words:[13)] *dis-*

12) We excluded function words such as articles, prepositions, and auxiliaries in our 50 MI3 collocations because we are interested in content words.

tress, rescue, search, safety, and *security* co-occurring with *maritime*; *accident, accidents, casualties, casualty, evacuation, incident, incidents, invasions, pollutants, pollution, prevention,* and *spills* co-occurring with *marine*. The second group concerns logistics: *transport* and *transportation* co-occurring with *maritime*; and *transport* co-occurring with *marine*. The third group is associated with industry: *industry labour, sector,* and *services* co-occurring with *maritime*; and *architects*, *industry, engineering, management, merchant*, *services,* and *shipbuilding* co-occurring with *marine*. These three groups showed *maritime* and *marine* collocates with different lexical items when sharing similar semantic meanings.

It is notable that there are two other groups displaying complementary distribution. One is an education-related group: *administration, bureau, committee, community, governmental*, *IMO*, and *organization* co-occurring with *maritime.* None of these words co-occurs with *marine*. The other is a bio-system and environment group: *biosafety, ecosystems, environmental, environment, evergreen*, and *mammal* co-occurring with *marine.* But none of these words co-occurs with *maritime*.

3.3.2. *ship* and *vessel*

Collocates of *ship* and *vessel* were also extracted by MI3 within four spans to the left and right. In Table 4 there are eight common collocates: *abandon, board, built, cargo, collision, contact, means,* and *new* in the second row and 42 different collocates in the third row.

Collocates of both *ship* and *vessel* are also related to several semantic fields. The first group includes accident type words: *abandon, arrested, collision, damage,* and *sank* co-occurring with *ship*; and *abandon, capsize, distress*, and *collision* co-occurring with *vessel*. The second group concerns the concept of building: *building, built, design,* and *construction* co-occurring with *ship*; and *built, capacity,* and *designed* co-occurring with *vessel*. The third type involves navigating operations-related terms: *position, operation(s), operators,* and *seaworthy* co-occurring with *ship*; and *overtake, overtaken, pass, escort, approaching, overtaking, crossing, propelled,* and *leeward* co-occurring with *vessel*.

13) Maritime English education has been required for both native and non-native crews because 70% of accidents at sea are related to human factors such as language misunderstanding (Chirea-Ungureanu and Georgescu 2009).

Table 4. Collocations of *ship* and *vessel* calculated by the MI3 test

(-4 ship +4)		(-4 vessel +4)	
ABANDON, BOARD, BUILT, CARGO, COLLISION, CONTACT, MEANS, NEW (8)			
ARRESTED	OPERATIONS	ALSO	KEEP
ASSIGNMENT	OPERATORS	APPROACHING	KEEPS
AVOIDANCE	PLAN	AVERAGE	LANE
BUILD	POSITION	AVOID	LEEWARD
BUILDING	PROVIDED	BEND	LESS
CERTIFICATE	RECYCLING	CALLING	NAME
COMPENSATION	REPORTING	CAPACITY	OVERTAKE
CONSTRUCTION	SANK	CAPSIZE	OVERTAKEN
CONTAINER	SEAWORTHY	CLEAR	OVERTAKING
DAMAGE	SECURITY	COMMAND	PASS
DESIGN	SERVICE	COMMERCIAL	POWER
DISTRIBUTION	SHORE	CONSTRAINED	PROPELLED
DOMAIN	SPECIFIC	CROSSING	PUSHING
EFFICIENCY	SURVEY	DESIGNED	RESTRICTED
EMPIRE	SYSTEM	DISTRESS	RULE
ENTITLED	TARGET	DRIVEN	SAFELY
EQUIPMENT	TRAINING	ENGAGED	SAFETY
FAMILIARISATION	TRANSFER	ESCORT	SEES
FLAG	UPRIGHT	EXPENSES	SMALL
FLY	WATER	FOLLOW	SPEED
GENERAL	(42)	GIVE	(42)
OPERATION		GROUPINGS	

3.4. Methodology for Visualization

Because we consider human spoken and written language as a system, language can be treated as a network system where the words are the nodes and their co-occurrences are linked nodes. In order to visualize collocations, we used two software tools: NetMiner Version 4 (Cyram 2013) and UCINET Version 6 (Borgatti et al. 2002) from social network analysis to discover nodes (called "source" in NetMiner) and links ("target" in NetMiner). NetMiner and UCINET are software tools for exploratory analysis and visualization of large network data based on Social Network Analysis to detect underlying patterns and structures of the network. They can be used for general research and teaching in social networks in various fields such as biology, economics, geography, information science, political science, and so on.

With the NetMiner tools, we compiled all bidirectional networks. We then used "spring embedding" algorithms to visualize our data. These algorithms computed by force-based graph layout algorithms have the

advantage of drawing networks clearly even though they take additional time (Lee 2012). When visualizing the networks, we showed the degree of thickness of linking lines using three divided scales depending on MI3 scores in order to better visualize the strength of collocations linking to each pair of keywords.

4. Visualization of Collocational Networks

4.1. Comparison of the Near-synonyms: *maritime* and *marine*

Two ego nodes meaning focal nodes, *maritime* and *marine*, are linked with their 50 alter nodes meaning not focal nodes but neighbors of some focal nodes in Figure 6.

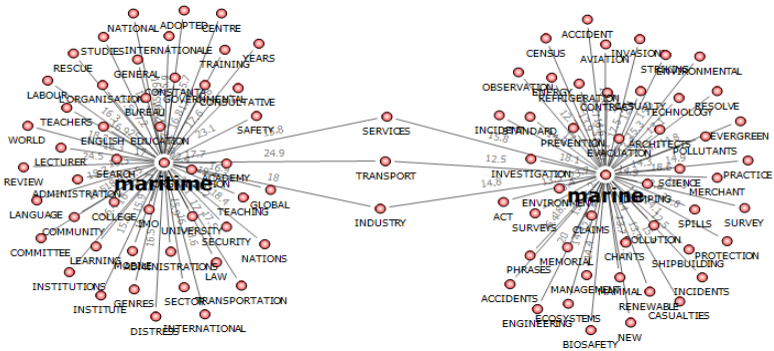

Figure 6. Network structure for near-synonyms: *maritime* and *marine.*

Among these nodes only three alter nodes (e.g., *industry, **services**,* and *transport*)[14] connect both *maritime* and *marine,* which means they are common collocations of both vocabularies. *Maritime* and *marine* connected within only one distance through these common alter nodes.[15]

14) An alter node, *services*, is in **BOLD** to show the greatest thickness of linking lines in three alter nodes. This strength between alter nodes and ego nodes will be discussed in detail in Section 4.3.

15) We would like to thank Professor Soosang Lee of Pusan National University for giving us his kind explanations and valuable comments. The distance means the geodesic length of shortest path between the two nodes.

4.2. Comparison of the Near-synonyms: *ship* and *vessel*

Following the same procedures, we produced the network graph in Figure 7. Two ego nodes, *ship* and *vessel,* are linked with their 50 alter nodes.

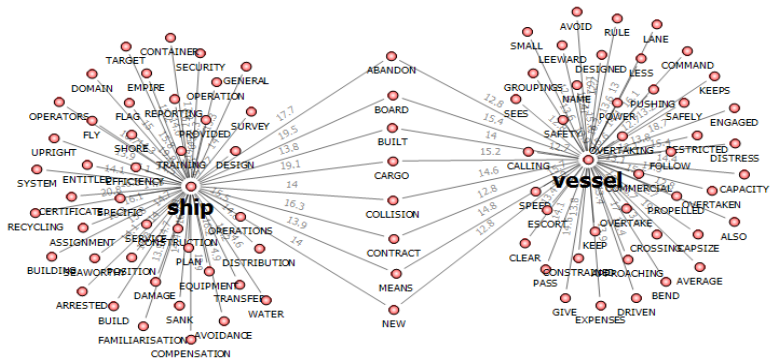

Figure 7. Network structure for near-synonyms: *ship* and *vessel.*

Among these nodes eight alter nodes (e.g., *abandon, board, built, cargo, collision, contact, means,* and *new*) connect both *ship* and *vessel,* meaning that they are common collocations of both vocabularies. S*hip* and *vessel* represent similarities because they are connected through these common alter nodes within only one distance.

4.3. Comparison of Networks to Link between Pairs of Near-synonyms

In the previous section we have examined collocates of *maritime-marine* and *ship-vessel* respectively. Now we will look at links between collocates of near-synonyms. UCINET was also employed to demonstrate collocational networks using the entire 200 nodes. Figure 8 shows *maritime-marine* and *ship-vessel* which are connected through several collocations.

In the network point of view the networks consist of one component, which demonstrates relational links between them without disconnection. Also, collocations which are linked to different ego nodes are called cut-off points or "cut-sets", if there is more than one node.

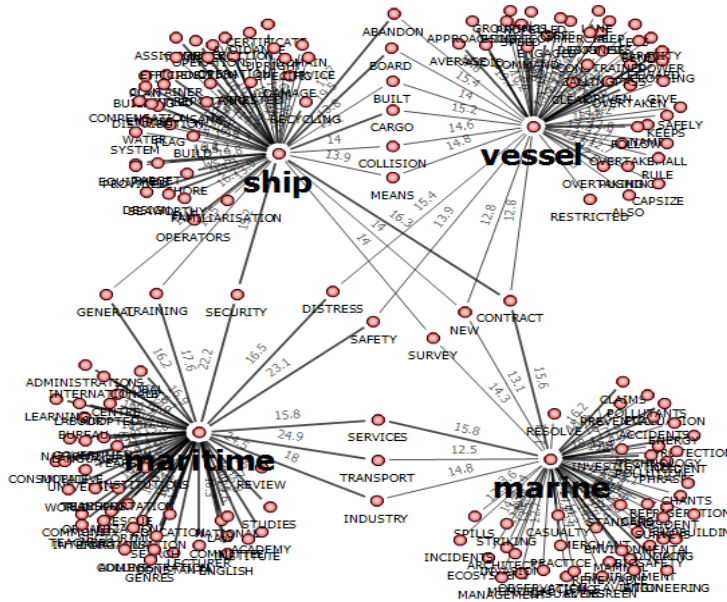

Figure 8. The near-synonyms network between *maritime-marine* and *ship-vessel.*

When we observed these connection patterns in terms of ego nodes, we found four new linked patterns which are interconnected by some cut-sets: (a) *maritime* and *ship* are connected by *general* (16.2:15.3)[16], ***security*** (22.2:19.2) and *training* (17.6:15.5), (b) *maritime* and *vessel* are connected by *distress* (16.5:15.4) and *safety* (23.1:13.9), (c) *marine* and *ship* are linked by *new* (13.1:14), ***contract*** (15.6:16.3), and *survey* (14.3: 14), and (d) *marine* and *vessel* are connected by *new* (13.1:12.8) and *contract* (15.6:12.8). Through all these cut-sets all four near-synonym groups can be connected. Because of these links any nodes in the network can be reached by other nodes. In all the cut-sets, lexical items such as *security* and *contract* connecting *maritime-ship* and *marine-ship,* respectively, show the greatest thickness of linking lines (depending on MI3 scores) if both scores are more than 15.5 and thereby visualize the strength of their collocations. The lexical item *services* connecting *maritime* and *marine* also shows the greatest thickness of

16) MI3 scores are represented in order of pairs of keywords in question. For instance, *general* (16.2:15.3) means that 16.2 is an MI3 score showing the strength between an ego node *maritime* and a cut-set collocate *general*, and 15.3 is that between an ego node *ship* and a collocate *general*.

linking lines in the cut-sets (e.g., *industry* (18:14.8), ***services*** (15.8:15.8), and *transport* (24.9:12.5)), as noted in sub-section 4.1.

The implication of these collocational networks is that we are able to use this network visualization to find connected words sharing a similar meaning (Baker 2006). What we found from our sample data is evidence for this. For example, collocates with the meaning of shipping industry (e.g., *industry, services,* and *transport*) are connected with a pair of keywords, *maritime* and *marine*. *Distress* and *safety* sharing the accidents meaning collocate with a pair of keywords, *maritime* and *vessel*. *Ship* and *vessel* are connected with the part of collocates sharing a meaning connected with accidents (e.g., *abandon* and *collision*). This suggests that we may be more likely to think of *maritime* and *marine* in terms of shipping or transporting activity and both *maritime* and *vessel* and *ship* and *vessel* in terms of dangerous navigational operations. It is notable that the connotation of *ship* and *vessel* includes unsafe and dangerous situations.

4.4. Comparison of 200 Node Networks and the Entire Network of All Collocations

It is hard at the moment to find network characteristics of all the collocations in our study because thousands of collocations were found. For this reason, we used statistical methods to extract significant collocations. However, it may be helpful to see how all the collocates construct their networks. We extended our study to all collocates to see how the language network is organized because we wanted to see if there were significantly different patterns of language networks when all nodes were connected. UCINET was employed to visualize the entire network of all collocations. We followed the same procedures for creating "spring embedding" in the previous section to produce these networks. As expected, the number of linked collocates was significantly increased.

The network structure of Figure 9 looks similar to that of Figure 8. In the case of large collocation groups, we found six common collocational groups between ego nodes: (a) *maritime* and *marine* (559), (b) *ship and vessel* (1,121), (c) *maritime* and *vessel* (667), (d) *maritime* and *ship* (871), (e) *marine* and *vessel* (573), and (f) *marine* and *ship* (573). Although we did not compare these two networks using statistical measures, we

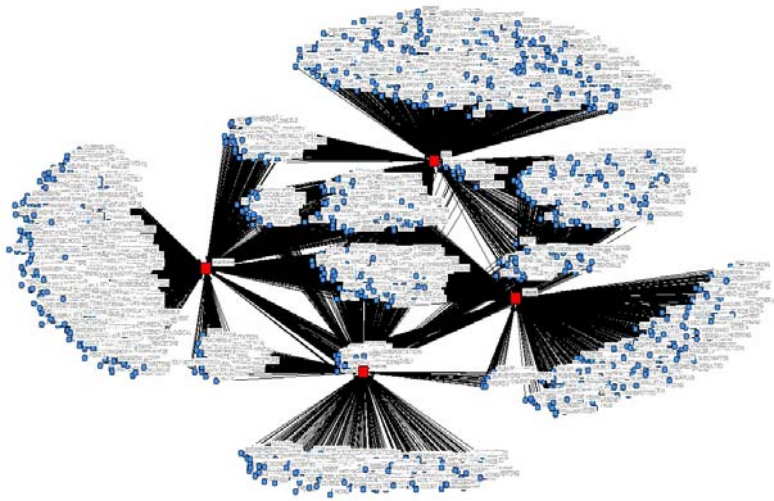

Figure 9. The network of all collocations of near-synonyms: *maritime-marine* and *ship-vessel.*

may assume that, as more nodes are added, similar patterns will appear. These results imply that our sampling method is an appropriate analysis approach.

From these results it seems that the networks discussed here show a "small world" effect, where every ego and alter node may be influential within the two distances. These findings can be compared with the research of Christakis and Fowler's (2010) three degrees of influence rule, where in the case of the "friend network" the scope of influence of friends extends to up to three distances: a friend, a friend of a friend, and a friend of a friend of a friend. What these results demonstrate is that human influence regarding political opinion, emotions, and recommendations is effective within the three distances. Therefore, it may be true that the all egos and alter nodes of near-synonyms have certain degrees of close relationships.

5. Conclusion

This paper has discussed the research field of language networks and conducted a case study in maritime vocabulary to visualize the semantic networks that the Maritime English corpus data has. We have

tried to visualize collocational language networks between highly frequent Maritime English near-synonyms: *maritime-marine* and *ship-vessel,* and to describe their structural characteristics. We found that visualization of these collocational networks enabled us to easily find connected words sharing a similar meaning by using the social network analysis tools: NetMiner and UCINET. We also found that the meaning of collocates of near-synonyms was grouped in several semantic fields. From our observation of collocational networks, we discovered that collocates with the meaning of shipping industry or transporting activity are connected with a pair of keywords, *maritime* and *marine*. Collocates sharing the meaning of accidents or dangerous navigational operations are connected with two pairs of keywords, *maritime-vessel* and *ship-vessel*. Thus, it is likely that the connotation of collocates co-occurring with *ship* and *vessel* includes unsafe and dangerous situations. Interestingly enough, in all the common collocates ("cut-sets"), lexical items such as *services, security,* and *contract* connecting *maritime-marine, maritime-ship,* and *marine-ship,* respectively, visually showed the greatest thickness of linking lines depending on MI3 scores. Finally we showed that the network structure of all collocations of near-synonyms looks similar to that of 200 sample collocations. In much larger collocation groups we also found six common collocational groups between ego nodes. This result indicates that our sample networks may reflect very similar characteristics for the entire network of all collocations.

As an example of language sampling of collocational networks, this study is a starting point for modeling language networks of Maritime English. The social network analysis method is useful for identifying collocational patterns and visualizing the structures and relationships of collocational networks between Maritime English keywords.

References

Barabasi, A.L. (2002). *Linked: The New Science of Networks*. Cambridge: Perseus.

Beavan, D. (2008). Glimpses Through the Clouds: Collocates in a New Light. *Proceedings of Digital Humanities*, University of Oulu.

Borgatti, S.P., Everett, M.G., and Freeman, L.C. (2002). *Ucinet for Windows: Software for Social Network Analysis.* Harvard, MA: Analytic Technologies.

Bocanegra-Valle, A. (2012). Maritime English. In Carol A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. UK: Wiley-Blackwell.

Chirea-Ungureanu, C. and Georgescu, M. (2009). Managing Cultural Diversity. *Proceedings of International Maritime English Conference* 21, 49-56.

Christakis, N.A. and Fowler, J.H. (2010). *Connected: The Surprising Power of our Social Networks and How They Shape our Lives*. New York: Little Brown and Company.

Church, K.W. and Hanks, P. (1990). Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics* 16.1, 22-29.

Church, K.W., Hanks, P., and Hindle, D. (1991). Using Statistics in Lexical Analysis. In U. Zernik (ed.). *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers. 115-164.

Culpeper, J. (2009). Words, Parts-of-speech and Semantic Categories in the Character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics* 14.1, 29-59.

Cyram (2013). *Netminer 4.0.* Seoul: Cyram Co. Ltd.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19.1, 61-74.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hong, S.C. and Jhang, S.E. (2010). The Compilation of a Maritime English Corpus for ESP Learners. *Korean Journal of English Language and Linguistics* 10.4, 963-985.

IMO. (2009). *Module Course 3.17: Maritime English*. Reading: IMO.

Jhang, S.E. and Byun, H.J. (2011). A Corpus-Based Lexical Analysis of Maritime English. *The New Korean Journal of English Language and Literature* 53.4, 247-266.

Jhang, S.E. and Parent K. (2011). The Vocabulary of Maritime English: Keyword Analyses of the English Homepages of Port Authorities Around the World. *Korean Journals of English Language and Linguistics* 11.4, 1065-1083.

Jhang, S.E. and Lee, S.M. (2013a). A Corpus-Based Lexical Analysis of Maritime English High School Textbooks. *Journal of Language Sciences* 20.1, 165-183.

Jhang, S.E. and Lee, S.M. (2013b). Clusters and Key Clusters in the Maritime English Corpus. *Journal of Language Sciences* 20.4, 199-219.

Jung, E.G. and Kang, B.M. (2011). A Network Analysis of Family Nouns. *The Linguistic Association of Korea Journal* 19.2, 209-235.

Lee, K.W. and Lee, J.Y. (2010). Lexical Network Analysis of Korean Dictionary. *Journal of Korealex* 16, 218-243.

Lee, S.S. (2012). *Network Analysis Methods* (Korean version). Seoul: Nonhyung.

Lewis, M. (1997). *Implementing the Lexical Approach: Putting Theory into Practice*. Hove, England: Language Teaching Publications.

Liang, W., Shi, Y., Tse, C.K., Liu, J, Wang, Y., and Cui, X. (2009). Compa-

rison of Co-occurrence Networks of the Chinese and English languages. *Physica A* 388, 4901-4909.

Manning, C.D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press.

Masucci, A. and Rodgers, G. (2006). Network Properties of Written Human Language, *Physical Review* E 74, 1-8.

McEnery, T. (2006). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. London: Routledge.

Nation, P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Newman, M. (2001). Scientific Collaboration Networks. Network Construction and Fundamental Results. *Physical Review* 64.

Oakes, M.P. (1998). *Statistics for Corpus Linguistics*, Edinburgh: Edinburgh University Press.

Rayson, P. (2003). *Matrix: A Statistical Method and Software Tool for Linguistics Analysis Through Corpus Comparison*. Ph.D. Thesis, Lancaster University.

Rayson, P. (2008). From Key Words to Key Semantic Domains. International *Journal of Corpus Linguistics* 13.4, 519-549.

Rayson, P. and Mariani, J. (2009). Visualising Corpus Linguistics. In M. Mahlberg, V. González-Díaz, and C. Smith (eds.), *Proceedings of the Corpus Linguistics Conference* Article 426. Liverpool, UK.

Schmitt. N. (2000). *Vocabulary in Language Teaching*. New York: Cambridge University Press.

Scott, J. (2000). Social Network Analysis: A Handbook. London: SAGE Publications.

Scott, M. (1997). PC Analysis of Key Words and Key Key Words. *System* 25.2, 233-245.

Scott, M. (2010). What Can Corpus Software Do? In A. O'keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge Handbooks, 136-151.

Scott, M. (2012). *WordSmith Tools Help Manual. Version 6*. Liverpool: Lexical Analysis Software.

Sinclair, J.M. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J.M., Jones, S., and Daley, R. (2004). *English Collocation Studies: The OSTI Report*. London: Continuum.

Stuart, K. and Botella, A. (2009). Corpus Linguistics, Network Analysis and Co-occurrence Matrices. *International Journal of English Studies*, Special Issue, 1-28.

Trenker, P. (2000). Maritime English: An Attempt at an Imperfect Definition. In *Second Asian IMLA Worshop on Maritime English (WOME 2A)*. Dalian, China: Dalian Maritime University, 1-8.

Walter, E. (2010). Using Corpora to Write Dictionaries In A. O'keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge Handbooks, 428-443.

Watts, D.J. and Strogatz, S.H. (1998). Collective Dynamics of Small-world Networks. *Nature* 393, 440-442.

Williams, G. (1998). Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles. *International Journal of Corpus Linguistics* 3.1, 151-171.

Zhou, S., Hu, G., Zhang, Z., and Guan, J. (2008). An Empirical Study of Chinese Language Networks. *Physica A* 387, 3039-3047.

Se-Eun Jhang
Korea Maritime and Ocean University
Department of English Language and Literature
727 Taejong-ro, Yeongdo-Gu, Busan 606-791, South Korea
E-mail: jhang@kmou.ac.kr

Sung-Min Lee
Korea Maritime and Ocean University
Department of English Language and Literature
727 Taejong-ro, Yeongdo-Gu, Busan 606-791, South Korea
E-mail: roy7942@hanmail.net

Received: October 31, 2013
Revised version received: November 29, 2013
Accepted: December 2, 2013