

해사영어 전문용어에 관한 연구

이성민*

*한국해양대학교 강사

요 약 : 본 연구에서는 해사영어어휘의 특징인 *ballast water*, *fore peak bulkhead*, *container*, *freight station charges*와 같은 n-gram의 복수 단어로 구성된 합성어 (multi-word compounds) 태깅(tagging)처리가 포함된 해사영어코퍼스를 구축하였다. 해사영어코퍼스는 백만 단어씩 수집한 학술, 법, 신문, 교과서 4개 하위 코퍼스로 구성된 총 400만 단어의 해사영어코퍼스로 구성되어 있다.

핵심용어 : 해사영어, 코퍼스언어학, 합성어, 자연언어처리, 태깅

Chapter 3. Maritime English Corpus

3.1 Introduction

- (1) Describing how to compile a maritime English corpus (MEC) including representativeness, balance, and size, stratified random sampling, web crawling and cleaning, and converting PDF to texts
- (2) Describing how to build a list of reference multi-word compounds
- (3) Presenting the procedures to tag multi-word compounds using customized Python coding
- (4) Comparing maritime English corpus with and without compounds based on basic statistics, word lists, n-gram lists, keyword lists, dispersion plots, and visualization

3.3 Corpus Compilation

- 3.3.1 Stratified Random Sampling
- 3.3.2 Web Crawling and Cleansing
- 3.3.3 Converting PDF to Texts

Stratified Random Sampling, Web Crawling and Cleansing, Converting PDFs to Texts



Chapter 3.

3.2 Corpus Design: Representativeness, Balance, and Size

Table 3.1 List of academic journal sources

Text ID	Title	Source	Date
A03	<i>Maritime Policy and Management</i>	http://www.tandfonline.com/doi/10.1080/03090940600570040	2010-2016
A04	<i>Journal for Maritime Research</i>	http://www.tandfonline.com/doi/10.1080/03090940600570040	2010-2016
A05	<i>Maritime Studies</i>	http://www.maritimestudiesjournal.com/	2012-2014
A06	<i>Oryzocopy and Navigation</i>	http://www.springer.com/engineering/mechanical-engineering/journal/13140	2010-2014
A07	<i>Aspen Review of the Law of the Sea and Maritime Law</i>	http://www.springer.com/law/international/journal/22100	2010-2013
A08	<i>ITN/Journal of Marine Affairs</i>	http://www.vmu.se/publications/vma-journal	2010-2016

3.4 Multi-word Compounds

Table 3.5 Types of general English compounds and maritime English compounds

	General English	Maritime English
a.	<i>ashtray, windmill, hotline</i>	<i>seafarer, shipyard, offshore, shipbuilding</i>
b.	<i>fast-food, icy-cold, call-girl</i>	<i>Hapag-Lloyd, double-hull, ABS-classed</i>
c.	<i>ice cream, income tax increase</i>	<i>coast guard, ballast water management</i>

† 교신저자 : roy7942@hanmail.net

No	All the vocabulary items	No	List of single words	No	List of multi-variant compounds
1	ab	1	ab	1	abast the beam
2	as	2	as	2	abandon ship
3	ast	3	ast	3	abandoned goods
4	abast	4	abast	4	abandonment
5	ab	5	ab	5	abandonment clause
6	aback	6	aback	6	abac analysis
7	abast	7	abast	7	abel brown
8	abast the beam	8	abandon	8	abel vector
9	abandon	9	abandonment	9	abac seaman
10	abandon ship	10	abandonment	10	abac seaman
11	abandoned goods	11	abast	11	abast ship
12	abandonment	12	abast	12	abast sea green
13	abandonment clause	13	abac seaman	13	abast water full
14	abastment	14	abac	14	abast flag
15	abac analysis	15	abac week	15	abast flag
16	abast	16	abac	16	abastee pennant
17	abel brown	17	abastion	17	abastee seaman
18	abel vector	18	abast	18	abastee bearing
19	abast	19	abastion	19	abastee contraband
20	abac seaman	20	abast	20	abastee pressure
...
18,790	col	0.001	0.001	0.001	colle price

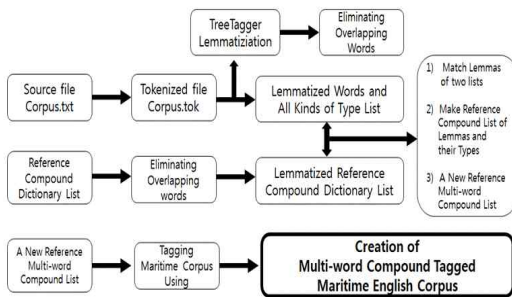
Table 3.9
All the vocabulary items from three different sources

3.6 Comparison of With and Without Compounds 3.6.1 Comparison of Basic Statistics

Table 3.10 Statistical results before tagging multi-word compounds

Genres	Token	Type	Type/Token Ratio (TTR)	Standardised TTR
Academy	989,943	79,639	8.47	56.17
News	997,835	39,790	4.12	55.03
Laws	985,621	15,500	1.63	30.05
Textbooks	989,610	53,617	5.64	54.11
Total	3,963,009	138,389	3.64	48.87

Fig. 3.6 Creation of multi-word compound tagged MEC



3.6 Comparison of With and Without Compounds 3.6.1 Comparison of Basic Statistics

Table 3.11 Statistics results after tagging multi-word compounds

Genres	Token	Type	Type/Token Ratio (TTR)	Standardised TTR
Academy	985,284	80,317	8.59	56.31
News	983,350	42,792	4.50	55.74
Laws	971,471	16,464	1.76	30.53
Textbooks	984,180	54,416	5.76	54.34
Total	3,924,285	142,041	3.77	49.29

Fig. 3.7 Result of compound tagged special vocabulary terms

en_route means that the ship is underway at sea on a course or courses, including deviation from the shortest direct route, which, as far as practicable for navigation purposes, will cause any discharge to be spread over as great an area of the sea as is reasonable and practicable. except as provided in paragraph 2 of this regulation, in ships delivered after 31 December 1979, as defined in regulation 1.28.2, of 4,000 gross tonnage and above other than oil tankers, and in oil tankers delivered after 31 December 1979, as defined in regulation 1.28.2, of 150 gross tonnage and above, no ballast water shall be carried in any oil fuel tank. where the need to carry large quantities of oil fuel render it necessary to carry ballast water which is not a clean ballast in any oil fuel tank, such ballast water shall be discharged to reception facilities or into the sea in compliance with regulation 15 of this annex using the equipment specified in regulation 14.2 of this annex, and an entry shall be made in the oil record book to this effect.

3.6 Comparison of With and Without Compounds 3.6.1 Comparison of Basic Statistics

3.6.2 Comparison of Word Lists, N-gram Lists, and Keyword Lists

Table 3.13 Comparison of untagged and tagged MEC word lists

No	Untagged Lex Corpus	Freq	No	Tagged Lex Corpus	Freq
1	TIRE	74189	1	TIRE	74000
2	OF	46678	2	OF	45993
3	Y	35384	3	Y	35380
4	AND	28234	4	AND	28232
5	TO	26234	5	TO	26232
6	IN	23881	6	IN	23869
7	BE	17869	7	BE	17869
8	A	17350	8	A	17344
9	OR	16344	9	OR	16497
10	SHALL	16380	10	SHALL	16380
174	WITHOUT	692	174	BALLAST_WATER	688
769	PACKED	165	769	BULK_CARGOES	177
1,777	CARTIDGE	54	1,777	CHIEF_ENGINEER	54
3,340	IGNATORY	21	3,340	MOULDED_DEPTH	27
3,863	LEAKAGES	15	3,863	ABANDON_SHIP	15
...
15,381	YEMEN	1	16,431	YEMENIA	1

3.6 Comparison of With and Without Compounds

3.6.1 Comparison of Basic Statistics

3.6.2 Comparison of Word Lists, N-gram Lists, and Keyword Lists

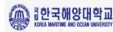


Table 3.15 Comparison of vocabulary on the keyword list

N	Untagged Law corpus	Keywords	N	Tagged Law corpus	Keywords
1	SHALL	35496.65	1	SHALL	35355.64
2	OR	15452.64	2	OR	155576.60
3	BE	10460.35	3	BE	10464.42
4	PARAGRAPH	4446.24	4	PARAGRAPH	4479.25
5	CONVENTION	4186.93	5	CONVENTION	4363.91
6	PROVISIONS	4127.62	6	PROVISIONS	4071.53
7	ANY	3826.46	7	ANY	3804.71
8	ACCORDANCE	3816.66	8	ACCORDANCE	3846.40
9	PROVIDED	3260.59	9	CERTIFICATE	3250.75
10	CERTIFICATE	3219.56	10	PROVIDED	3259.50
11	SPACES	3019.47	11	SPACES	3174.75
34	COMPETENT	1849.52	34	DANGEROUS GOODS	1844.39
69	OTHER	1041.42	69	BALLAST WATER	1048.02
424	VOYAGES	219.70	424	LIQUEFIED GASES	222.15
-	-	-	-	-	-
1,968	ETHYLENE	24.04	2,183	RATING	24.05
1,969	ABOARD	-23.92	2,184	ELEMENT	-23.53
-	-	-	-	-	-
2,759	WAS	-3444.55	3,070	WAS	-3422.31

3.6 Comparison of With and Without Compounds

3.6.1 Comparison of Basic Statistics

3.6.2 Comparison of Word Lists, N-gram Lists, and Keyword Lists



Table 3.14 Comparison of the untagged and tagged MEC 4-grams

N	Untagged Law Corpus	Freq	N	Tagged Law Corpus	Freq
1	BY ACCORDANCE WITH THE	803	1	BY ACCORDANCE WITH THE	807
2	WITH THE REQUIREMENTS OF	401	2	FOR THE PURPOSE OF	188
3	FOR THE PURPOSE OF	188	3	THE PROVISIONS OF THIS	174
4	THE PROVISIONS OF THIS	174	4	THE DATE ON WHICH	158
5	WITH THE PROVISIONS OF	162	5	CONSTRUCTED ON OR AFTER	118
6	THE DATE ON WHICH	121	6	IN THE CASE OF	107
7	CONSTRUCTED ON OR AFTER	111	7	WITH THE PROVISIONS OF	269
8	IN THE CASE OF	106	8	BY THE COMPETENT AUTHORITY	260
9	BY THE COMPETENT AUTHORITY	260	9	WITH THE REQUIREMENTS OF	269
10	TO THE PROVISIONS OF	268	10	ADOPTED BY THE	262
11	ADOPTED BY THE ORGANIZATION	262	-	ORGANIZATION	-
-	-	-	18	TO THE PROVISIONS OF	78
271	TAKE INTO ACCOUNT THE	10	271	BALLAST WATER AND	10
1,859	A MEMBER OF THE	10	1,859	MEMBERS	10
4,577	CANAL OF VENTILATION A	11	4,577	BEYOND PROHIBITED AREA	11
13,454	REQUIREMENTS OF A VESSEL	6	13,454	OF OTHER PROHIBITED AREAS	6
10,816	THE HULLING HAS ATTACHED	5	10,816	AND THE PORT/STARBOARD	5
-	-	-	-	STORAGE	-
20,719	ZONE IN AT AREA	5	20,660	THE HULLING HAS ATTACHED	5