



## 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

**Thesis for the Degree of Doctor of Philosophy**

**Syntactic and Semantic Patterns of Domain-specific Multiword Units  
in Marine Accident Investigation Reports**

by

Yilian Qi

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

in the Department of  
English Language and Literature

Korea Maritime and Ocean University

January 2019



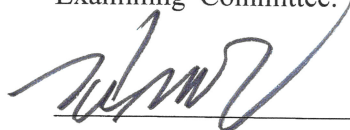
# Approval

Name: Yilian Qi

Degree: Doctor of Philosophy

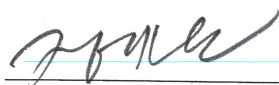
Title of Thesis: Syntactic and Semantic Patterns of Domain-specific Multi-word Units  
in Marine Accident Investigation Reports

Examining Committee:



Dr. Jae-Hoon Kim

Chair, Professor in Computer and Information Engineering  
Korea Maritime and Ocean University



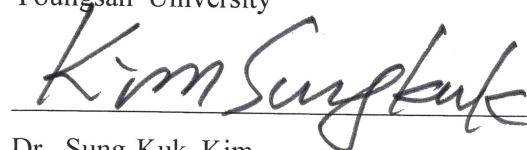
Dr. Se-Eun Jhang

Senior Supervisor, Professor in English Linguistics  
Korea Maritime and Ocean University



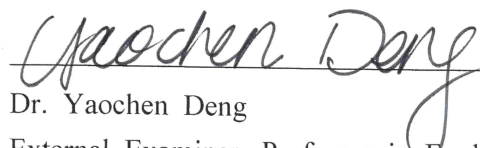
Dr. Shin Ho Kim

Professor in English Linguistics  
Yongsan University



Dr. Sung-Kuk Kim

Adjunct Professor in International Maritime Transportation Science  
Mokpo National Maritime University



Dr. Yaochen Deng

External Examiner, Professor in English Linguistics  
Dalian University of Foreign Languages, China

Data Approved: December 14, 2018





# **Syntactic and Semantic Patterns of Domain-specific Multiword Units in Marine Accident Investigation Reports**

Yilian Qi

Department of English Language and Literature  
*Graduate School of Korea Maritime and Ocean University*

## **Abstract**

The present study is a systematic corpus-based investigation of the domain-specific multiword units (henceforth MWUs) in marine accident investigation reports (henceforth MAIR), with a view to characterizing their most prominent syntactic, semantic and functional features.

To achieve these principal objectives, the target MWUs were first identified by applying a new approach, which incorporates the notion of ‘meaning’ into statistical-based measures. This method ensures the domain-specific MWU extraction to the largest extent and provides valid data for the subsequent analysis. Through proposing a three-dimensional analytical framework, this study has obtained the following findings:

First, the domain-specific MWUs are largely composed of two-word sequences, while the occurrences of 4- and 5-word MWUs are relatively rare. Among all the target MWUs, only 1.10% of the expressions occur very commonly within the genre (>1,000 times). By contrast, the majority of the expressions (70.97%) occur with the frequency less than 100 times. The skewed distribution indicates that MAIR genre tends to employ a wide variety of domain-specific MWUs rather than repetition of a small number of common expressions.

Second, in terms of the syntactic features of the domain-specific MWUs, NP structure is the most commonly employed grammatical type. The abundant use of this structure implies that the domain-specific meaning of MAIR genre is largely carried

in the nominal group. Apart from NP structure, there is also a marked prevalence of VP structures among the domain-specific MWUs in MAIR genre and these MWUs present structural variation. Of all the VP-based patterns, the ‘verb phrase with active verb’ pattern stands out since it incorporates a large number of action verbs, which are used to describe the actions done by people. The wide use of these phrases implies that MAIR genre tends to highlight the people’s roles during the accidents, with particular attention to the information about what or who caused or performed the activity. Similarly, PP structures were also frequently adopted by the domain-specific MWUs, especially the pattern beginning with preposition *of*. This pattern was mostly used to specify possessions. It thus can be inferred that the information that provided in MAIR genre tends to be concrete and specific.

Third, by conducting a functional analysis of the target MWUs, it was found that the primary function of the domain-specific MWUs is to express referential meanings and contribute to the thematic development. Furthermore, due to their multifunctional nature, some referential MWUs also perform the function of stance and discourse organizing. When expressing stance, most MWUs express impersonal epistemic stance, with the purpose of minimizing the imposition of the reporters’ opinions. Other word sequences appear to be deontic in nature, as they are mainly realized by the MWUs incorporating with *require* or modal verbs. The primary function of these MWUs is to set out the obligations and issue suggestions for the agents according to certain norms and regulations. When functioning as discourse organizer, the domain-specific MWUs usually adopt the pattern of ‘*that*-clause controlled by main verbs in active voice’ to introduce the topics. Unlikely, when using for elaborating the topics, they tend to clarify the logical relationships, especially the causative-resultative relation, rather than providing additional information in MAIR genre.

Fourth, the distinctive semantic features of the domain-specific MWUs can be best reflected when these MWUs perform the functions of activity identification and specification. For instance, most domain-specific MWUs used for describing activities are of general nature, but they convey specialized meaning in MAIR genre.

Similarly, when domain-specific MWUs are used to provide tangible or intangible frames for specifying certain attributes, the use of these MWUs in MAIR genre is significantly deviant from their use in general English register.

In all, by gaining insights into the salient features of the domain-specific MWUs in MAIR genre, the present study may make contributions and implications in the following aspects: the construction of extraction method for domain-specific MWUs, the compilation of maritime-specific MWU list, the teaching and learning of maritime English, especially the maritime-specific MWUs, and providing reference for writing MAIR to the experts who are from non-native English speaking countries.



## Acknowledgement

I would like to express my sincere appreciation to all those who have contributed to making this thesis a reality. First and foremost, I am forever indebted to my supervisor, Professor Se-Eun Jhang for professional guidance, thought-provoking comments and intellectual stimulation throughout the research process. Without his illuminating instruction, patience and encouragement, this thesis would not have come to fruition.

Special thanks also go to the rest of my thesis committee: Professor Jae-Hoon Kim, Professor Shin-Ho Kim, Professor Sung-Kuk Kim, and Professor Yaochen Deng. Their insightful suggestions and enlightening questions on an earlier manuscript incentivized me to broaden my research through various perspectives.

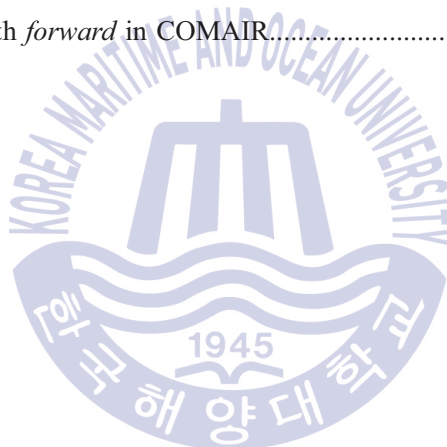
A deep sense of gratitude goes to my husband Zhi Liu and my son Zirui Liu for their continuous support and encouragement during the three-year study. It is they who have walked me through all the laughter and tears. My achievement owes a great deal to my beloved family. I am also grateful to my parents, who have sacrificed too much for my education, and my mother-in-law for her care and love.

Last but not the least, I would like to express my gratitude to my colleagues from School of Foreign Languages, Dalian Maritime University. They kindly shared my heavy teaching tasks during my stay in Korea Maritime and Ocean University. Any progress that I have made is the result of their profound concern and selfless devotion.

## List of Tables

Table 3.1 Structural taxonomy designed for the present study.....	39
Table 3.2 Functional categories designed for the present study .....	43
Table 3.3 Descriptive statistics for COMAIR.....	46
Table 3.4 Overall statistics of COMAIR .....	46
Table 3.5 Descriptive data of BNC Baby.....	48
Table 4.1 Procedures of domain-specific MWU identification.....	55
Table 4.2 Distribution of the n-grams of varied lengths (step 1).....	56
Table 4.3 A list of Keywords obtained by referencing COMAIR against BNC Baby (step 2)..	56
Table 4.4 Distribution of the keyword-grams of varied lengths (step 3).....	57
Table 4.5 A 2-by-2 Contingency Table.....	62
Table 4.6 Distribution of the keyword-grams above LL threshold (LL score $\geq$ 3.84) (step 4).....	64
Table 4.7 First 10 keyword 2-grams arranged in the descending order of LL value.....	65
Table 4.8 First 10 keyword 3-grams arranged in the descending order of LL value.....	65
Table 4.9 First 10 keyword 4-grams arranged in the descending order of LL value.....	66
Table 4.10 First 10 keyword 5-grams arranged in the descending order of LL value.....	66
Table 4.11 Distribution of the domain-specific MWUs of differing lengths (step 5).....	71
Table 4.12 A list of domain-specific MWUs (frequency $\geq$ 40).....	71
Table 5.1 Distribution of the domain-specific MWUs of differing lengths (frequency $\geq$ 40).....	73
Table 5.2 Distribution of the domain-specific MWUs across frequency bands (frequency $\geq$ 40).....	75
Table 5.3 A list of domain-specific MWUs occurring over 1000 times in COMAIR.....	76
Table 5.4 Structural classifications of the domain-specific MWUs in COMAIR (frequency $\geq$ 40).....	77
Table 6.1 Frequency distributions across primary discourse functions (frequency $\geq$ 40).....	81
Table 6.2 Frequency distributions across subcategories of stance function (frequency $\geq$ 40).....	85
Table 6.3 Adjectives used in the pattern of ‘copula <i>be</i> + adj.’ .....	86
Table 6.4 A list of stance MWUs arranged in a descending order (frequency $\geq$ 40).....	87

Table 6.5 A list of discourse organizing MWUs arranged in a descending order (frequency $\geq$ 40).....	91
Table 6.6 Frequency distributions across subcategories of discourse organizing Function (frequency $\geq$ 40).....	91
Table 6.7 Top 10 most frequent topic introduction MWUs in COMAIR.....	92
Table 6.8 Frequency distributions across functional subcategories of Referential MWUs (frequency $\geq$ 40).....	98
Table 6.9 Repetition rate of functional category based on the token/type ratio.....	99
Table 6.10 Most common MWUs identifying vessels in COMAIR.....	101
Table 6.11 Domain-specific MWUs containing <i>secure</i> in COMAIR.....	102
Table 6.12 MWUs combining <i>heading</i> in COMAIR.....	109
Table 6.13 MWUs with <i>forward</i> in COMAIR.....	111



## List of Figures

Figure 3.1 Snapshot of the website where the data was downloaded.....	45
Figure 4.1 First 20 concordance lines of <i>a passage</i> in COMAIR.....	70
Figure 5.1 Types of domain-specific WMUs of various lengths in COMAIR.....	73
Figure 5.2 Tokens of domain-specific MWUs of various lengths in COMAIR....	74
Figure 5.3 Distribution of structural types.....	77
Figure 5.4 Distribution of structural tokens.....	78
Figure 6.1 Distribution of functional types.....	81
Figure 6.2 Distribution of functional tokens.....	81
Figure 6.3 Distributions of the syntactic patterns serving the function of epistemic stance.....	86
Figure 6.4 Snapshot of the concordance lines of <i>shall be</i> in COMAIR.....	89
Figure 6.5 Snapshot of the concordance lines of <i>shall be</i> in BNC Baby.....	89
Figure 6.6 Snapshot of the concordance lines of <i>established that</i> in COMAIR...	92
Figure 6.7 Snapshot of the concordance lines of <i>established that</i> in BNC Baby..	93
Figure 6.8 Snapshot of the MWUs incorporating <i>prompted</i> in COMAIR.....	94
Figure 6.9 Snapshot of the MWUs incorporating <i>prompted</i> in BNC Baby.....	94
Figure 6.10 Snapshot of the concordance lines of <i>associated with</i> in COMAIR..	96
Figure 6.11 Snapshot of the concordance lines of <i>associated with</i> in BNC Baby.	96
Figure 6.12 Distribution across functional categories of Referential MWUs (types)....	98
Figure 6.13 Distribution across functional categories of Referential MWUs (tokens)..	98
Figure 6.14 Snapshot of the concordance lines of <i>to secure the</i> in BNC Baby...	103
Figure 6.15 Snapshot of the concordance lines of <i>appreciation of</i> in BNC Baby.	108
Figure 6.16 Snapshot of the concordance lines of <i>appreciation of</i> in COMAIR..	108
Figure 6.17 Snapshot of the MWUs incorporating <i>heading</i> in BNC Baby.....	109
Figure 6.18 Snapshot of the concordance lines of <i>a heading of</i> in COMAIR.....	110
Figure 6.19 Snapshot of the MWUs incorporating <i>forward</i> in BNC Baby.....	112



## Contents

<b>Abstract.....</b>	<b>i</b>
<b>List of Tables.....</b>	<b>v</b>
<b>List of Figures .....</b>	<b>vii</b>
<b>Chapter 1 Introduction .....</b>	<b>1</b>
1.1. Background of this study .....	1
1.2. Objectives of this study.....	3
1.3. Significance of this study.....	4
1.4. Terminological issues .....	5
1.5. Organization of this dissertation .....	6
<b>Chapter 2 Theoretical background .....</b>	<b>8</b>
2.1. Understanding the notions of phraseology .....	8
2.2.1. An overview of influential notions of phraseology .....	9
2.1.2. Parameters of defining MWUs .....	13
2.1.3. Operational definition of MWUs .....	17
2.1.4. An overview of influential taxonomy of phraseology .....	19
2.2. Theoretical discussion of MWUs .....	23
2.2.1. Theoretical framework of this study .....	23
2.2.2. Nature of multiword units .....	25
2.2.3. Previous studies of phraseology .....	29
<b>Chapter 3 Analytical framework and research design .....</b>	<b>37</b>
3.1. Analytical framework .....	37
3.1.1 Analytical framework for syntactic features of domain-specific MWUs.....	38
3.1.2. Analytical framework for semantic features of domain-specific MWUs .....	40

3.1.3. Analytical framework for functional features of domain-specific MWUs .....	42
3.2. Research questions.....	43
3.3. Corpora used in this study .....	44
3.3.1. Corpus of Marine Accident Investigation Reports (COMAIR).....	44
3.3.2. British National Corpus Baby (BNC Baby).....	47
3.4. Tools and procedures for data analysis .....	48
3.4.1. Tools for data processing .....	48
3.4.2. Procedures for data analysis .....	49
3.4.3. Inter-rater reliability .....	50
3.5. Summary.....	51
<b>Chapter 4 Identification of domain-specific MWUs in the COMAIR .....</b>	<b>52</b>
4.1. Current approaches to MWU extraction .....	52
4.2. My proposed approach to domain-specific MWU extraction .....	53
4.3. The detailed process of domain-specific MWU extraction .....	55
4.3.1. Step 1: N-gram retrieval .....	55
4.3.2. Step 2: Keyword-gram extraction .....	56
4.3.3. Step 3: Measuring the association strength of keyword-grams.....	58
4.3.4. Step 4: Filtering out process .....	66
4.3.5. Step 5: Domain-specific MWU identification .....	70
<b>Chapter 5 Frequency distributions and syntactic features of domain-specific MWUs.....</b>	<b>72</b>
5.1. Frequency distributions of domain-specific MWUs .....	72
5.1.1. Frequency distributions of domain-specific MWUs in various lengths .....	72
5.1.2. Overall frequency distribution across different frequency bands.....	74
5.2. Syntactic features of domain-specific MWUs .....	76

<b>Chapter 6 Functional and semantic features of domain-specific MWUs .....</b>	<b>80</b>
6.1. Distributions across primary discourse functions .....	80
6.2. Multiple functioning .....	82
6.3. Stance MWUs .....	84
6.3.1. Notion of stance MWUs .....	84
6.3.2. Stance MWUs in COMAIR .....	84
6.4. Discourse organizing MWUs .....	90
6.4.1. Notion of discourse organizing MWUs .....	90
6.4.2. Discourse organizing MWUs in COMAIR .....	90
6.5. Referential MWUs .....	96
6.5.1. Notion of referential MWUs .....	97
6.5.2. Referential MWUs in COMAIR .....	97
6.6. Summary .....	112
<b>Chapter 7 Conclusions and implications .....</b>	<b>113</b>
7.1. Summary of the major findings .....	113
7.2. Implications of this study .....	116
7.3. Limitations of this study .....	117
<b>References .....</b>	<b>118</b>
<b>Appendix .....</b>	<b>132</b>

# Chapter 1

## Introduction

### 1.1. Background of this study

Over the last few decades, the issue of phraseology has taken center stage in many fields of linguistics. As stated by Ellis, “phraseology pervades theoretical, empirical, and applied linguistics. Like blood in systemic circulation, it flows through heart and periphery, nourishing all” (Ellis 2008: 9). Although it sounds radical, this view reflects the central status of phraseology among linguists’ concerns. One reason behind such popularity lies in the ubiquitous use of phraseology in natural language, which has been found to take a large proportion of natural language production. Apart from this, phraseology has also received much attention for the crucial role it plays in meaning creation and language description. According to Sinclair, “the normal primary carrier of meaning is the phrase not the word; the word is the limiting case of the phrase, and has no other status in the description of meaning.” (Sinclair 2004:148; Sinclair 2008a: 409) He then indicates a clear superiority of phraseological approach to the one focusing on the isolated words by arguing that “one of the great strengths of a phraseological approach is the preservation of the integrity of text for much longer than alternative approaches to description, and in turn this entails the preservation of meaning” (Sinclair 2008b: xvii).

Undoubtedly, the widespread concern of this issue brings about many landmark studies, including Altenberg (1998), Biber (1999, 2004, 2006), Granger and Meunier (2008), Hunston (2000, 2008), Moon (1997, 1998), Sinclair (1991, 2004, 2006), Tognini-Bonelli (2001), and many others. To be specific, Sinclair (1991) has proposed the famous “idiom principle” and the notion of “extended unit of meaning”. In Renouf and Sinclair’s study (1991), they have highlighted the important roles of collocational

frameworks. Altenberg (1998) has pinpointed the prevalence of recurrent word combinations along with their formal flexibility and pragmatic conventionality. Biber (2006) has opened up a new avenue of research by discussing the important features of lexical bundles across spoken and written registers of university language. Hunston (2008) further develops the theory of pattern and meaning, with a notion of “semantic sequence”.

Inspired by the above studies and their related ideas, an overwhelming amount of research follows this line to explore different phraseological patterns and their ways to realize meaning in various subsets of language. Despite the far-reaching significances they bring to this field, the selected types of language that most phraseological studies target at are restricted to general English or academic texts, whereas systematic studies of phraseology in ESP field are few and far in between.

It is well known that there is a great deal of variation across register (Biber 1988, Biber, Johansson, Leech, Conrad, and Finegan 1999, Hofland and Johansson 1982). In linguistic enquiry, the texts, which are written for professional purposes with a narrower scope of target readers (usually professionals in specialized fields), are often referred to as ESP. These specialized texts differ from general English texts in that they present stylistic varieties with particular features: unique discourse organizational patterns, highly professional content, narrower lexical range, a large number of (semi-) technical notions, high frequencies of binomials, etc. All these features contribute to make the phraseology of ESP, especially the domain-specific phraseological patterns, different to a wide extent from that of general English. Therefore, deeper knowledge of their behavior is particularly promising. It not only determines how meaning is created in this type of language that shows a specialized grammar and vocabulary. More importantly, it is likely to afford insights into the nature of specialized languages and into the way specialized genres function. Therein lies the intended contribution of the current research.

## 1.2. Objectives of this study

The current research focuses on the restricted language of marine accident investigation reports (MAIR) in the field of maritime domain. The MAIR is one of the most essential written text types among all the maritime-related writings, since safety issues are always one of the greatest concerns in maritime domain. To prevent and avoid marine accidents, the MAIR is required to be provided in each accident investigation. Overall, it functions as a platform where experts can report investigation findings, explain the causes of the accident and express recommendations for other vessels.

By systematically exploring the frequently recurring domain-specific word combinations used in this particular text type, this study intends to shed light on the salient features of these phrases, along with their meanings and functions. Specifically, four principal objectives are considered in detail.

One general aim of work on routine and idiomatic language use is to identify the main recurrent phrasal constructions in a certain text and explain why they are frequent (Stubbs 2005). Hence, the first objective of the present study is to reveal the distinctive features of the domain-specific multiword sequences used in MAIR genre.

The second objective lies in proposing an efficient approach to extracting domain-specific word sequences from specialized corpus. Currently, the widely used statistical methods for phrase extraction are either frequency-based or ‘association measures’, each of which has come under criticism for not achieving high precision and efficiency in phrase extraction, particularly in extracting the domain-specific phrases. To improve extraction performance, this study attempts to propose a new approach by incorporating the notion of ‘meaning’ into statistical-based measure, since the domain-specific word sequences are usually the conveyer of the meaning in specialized texts. It is thus believed that the ‘meaning of text’ can be used as an efficient linguistic filter which allows us to extract the phrases of this type.

The third objective is to explore the quantitative features of domain-specific word sequences, including the frequency and distribution of varied types of phrases.

The last objective of this study is to investigate the domain-specific phrases from the perspectives of form, meaning and function. To achieve this goal, some concerns are specifically tackled: the extent to which structural patterns contribute to the overall meaning; the functions that these phrases tend to perform in the text; and the ways in which these patterns, meanings and functions are co-selected.

### 1.3. Significance of this study

Compared with similar studies conducted previously, this empirical investigation of the domain-specific phrases in the MAIR is not just significant theoretically and methodologically. It also offers some pedagogical implications as described below.

Traditionally, the research on phraseology has paid much attention to the ‘classical’ idioms, whose meaning cannot be predicted from those of their component words (e.g., *kick the bucket*). However, the multitude of findings in recent studies has shown that the phraseological units in natural language are not only comprised of idioms. In fact, there exists a much larger set of non-compositional multiword units which occur more frequently in natural discourse than idioms, such as conventionalized expressions and semi-fixed patterns (Moon 1998: 63-64). The present study therefore extends the scope of phraseology research by investigating the domain-specific phrases which apparently fall outside the limits of traditional idioms. Corpus evidence shows that the multiword sequences of this type not only include collocations, but also lexicalized sentence stems and clause constituents, which have not been thoroughly investigated before. By doing so, we hope to gain an overall understanding of the salient features of this type of phraseological patterns in MAIR genre and meanwhile, provide more insights into and mechanism of meaning realization in language use.

Methodologically, the present study is innovative in proposing a hybrid method for extracting the domain-specific multiword sequences. First, we refined the current statistical-based measures, since they have been recognized not producing satisfactory results. For example, the frequency threshold method neglects the internal association



of the ‘co-occurring’ words, which results in a considerable number of ‘phraseologically uninteresting’ sequences retrieved (e.g., *in the, of a, to a*) (Altenberg 1990: 133). Although ‘association measures’ can help determine whether the co-occurring words in sequences are meaningfully associated or not, they are typically restricted to the association within 2-word sequences. As for sequences longer than 2 words, frequency-based method is usually applied with a higher priority than the association measure. Then, we incorporate the notion of ‘meaning’ into the refined statistics-based approach. The proposed method is proved to achieve high precision and efficiency in identifying the domain-specific multiword sequences used in the specialized corpus.

Pedagogically, the investigation of the use of domain-specific phrases in ESP fields can serve as a starting point for learning and teaching practice. Swales (1990) points out that every genre of EAP and ESP has its own phraseology, and learning to be effective in the genre involves learning this. Nattinger and De Carrico (1992) also argue for the lexical phrase as the pedagogically applicable unit of pre-fabricated language, “for a great deal of the time anyway, language production consists of piecing together the ready-made units appropriate for a particular situation and... comprehension relies on knowing which of these patterns to predict in these situations. Our teaching therefore would center on these patterns and the ways they can be pieced together, along with the ways they vary and the situations in which they occur.” (Nattinger 1980: 341). The present study will describe the recurrent domain-specific phrases that may not be detectable by personal intuition, and identify their salient patterns and functions. By gaining insights into the formulaic nature of MAIR discourse, it is hoped to provide valuable reference for writing MAIR to the experts who are from non-native English speaking countries.

#### **1.4. Terminological issues**

In previous studies, the notion of phraseology is defined from different perspectives and under various sub-disciplines of linguistics. There is thus a plethora of concepts



describing the phenomenon. For instance, terms that have been used in corpus-based or corpus-driven studies include “lexical bundles” (Biber, Conrad, and Cortes 2003), “recurrent word combinations” (Altenberg 1998), “phraseological units” (Granger 2008), “multi-word units” (Granger and Paquot 2008; Sinclair 2008), “phraseme” (Granger and Paquot 2008), “multi-word expressions” (Sinclair 2008), “extended units of meaning” (Sinclair 2004), “phraseological sequences” (Wei 2009), etc. While “n-grams” (Manning and Schutze 2000; Jurafsky and Martin 2009) and “clustering” (Manning and Schutze 2000) are the terms proposed in the field of natural language processing. Other terms have been brought forward either from a pedagogical point of view, like “lexical phrases” (Nattinger and De Carrico 1992), “lexical chunks” (Ellis 1996), “recurrent sequences” (De Cock 2003), or from a psycholinguistic perspective, such as “formulaic language” (Wray 2000). The diversity of terms and definitions reflects the heterogeneous nature of phraseology that word combinations come in many different shapes and forms. As Howarth (1998: 25) points out, “such terms may be used too loosely as labels for a wide range of phenomena that may, under closer examination, differ significantly from each other.”

Based on the above, an operational definition of phraseology has to be provided in the current research to clarify which type of phrases is being investigated. In this study, we employed the term ‘multi-word unit (MWU)’ in a non-technical and broad sense to cover various types of lexico-grammatical sequences that frequently occur in the MAIR. Among these types, this study specifically focuses on the domain-specific MWUs as the target of the research. The detailed defining parameters, together with the linguistic criteria for automatic identification will be elaborated in Section 2.1.

### **1.5. Organization of this dissertation**

In response to the research objectives, this thesis developed into seven chapters. Following on from this introductory chapter, Chapter 2 mainly addresses the theoretical issues concerning phraseology, including an overview of some influential notions of phraseology, major theoretical frameworks and classifications, and the

previous studies conducted relating to this issue. Chapter 3 deals with the methodological issues. It first introduces the analytical framework specifically developed for the current empirical investigation. It is then followed by the research questions to be addressed. Afterwards, the applied methodology is discussed, along with the presentation of the corpora as well as the tools and procedures for data analysis. In Chapter 4, the proposed method for extraction of domain-specific MWUs from the study corpus is introduced. Chapter 5 and Chapter 6 form the core part of this dissertation. They discuss the distinctive features of phrases used in the MAIR in detail from the perspectives of structure, meaning and function. Finally, the results are discussed and the features are generalized. Chapter 7 brings together the main findings of the research, discusses the limitations and implications for future studies.



## Chapter 2

### Theoretical Background

This chapter is dedicated to a brief overview of the area of phraseology study, with particular emphasis on introducing some major terminologies in phraseology and reviewing the relevant previous studies in this field. Section 2.1 first lists some influential notions about phraseology. Based on the above discussion, it then proposes an operational definition of MWUs for the present study. Section 2.2 introduces some major theoretical frameworks in this field, through which the status and nature of MWUs are displayed and the widely adopted taxonomy for classifying MWUs is explicated. This section concludes with selecting a taxonomy which will be adopted for the detailed analysis in the present study. Section 2.3 mainly reviews the phraseological studies conducted under corpus-driven and corpus-based paradigm, even though some influential theoretical-driven studies are mentioned at the beginning of the section. In general, all of discussions above are intended to lay a theoretical foundation for the empirical part of the present study.

#### 2.1. Understanding the notions of phraseology

While the notion of phraseology is a very widespread concept, due to the heterogeneous nature of MWUs and the lack of a coherent theoretical and empirical model, developing a comprehensive definition of the phenomenon still remains one of the foremost problems in the area (Schmitt and Carter 2004: 2). In fact, criteria used for definition and classification vary depending on the research interest and there exist considerable overlap of assumptions, concepts, and findings less transparent than one would like (Gries 2008). Therefore, for the validity of research, it is of significance to make it explicit what kinds of sequences are identified as “MWUs” in the present study. This section reviews some influential terminologies in phraseology, based on which four defining parameters are arrived at and an operational definition of MWUs

is proposed for the present research.

### **2.1.1. An overview of influential notions of phraseology**

The notion of phraseology has been defined from various perspectives under linguistic discipline. Due to the diversity in assumptions and research methods, there is no single satisfactory definition of phraseology at present. Among all the definitions, it is of significance to discuss the following approaches, since they have been widely regarded as the milestones in the field of phraseological studies. They are the Russian tradition of phraseology, the approaches of psycholinguistics, cognitive grammar, and corpus-driven approach.

#### **2.1.1.1. The Russian tradition of phraseology**

The traditional approach to phraseology derives from the linguists from the former Soviet Union and other Eastern European countries, such as Russian scholars like Vinogradov (1947) and Amosova (1963). The foundation of the early Russian scheme is to regard phraseology as a continuum along which word combinations are located, with the most opaque and fixed ones at one end and the most transparent and variable ones at the other. To be specific, this continuum includes a specific subset of linguistically defined expressions such as idioms, proverbs, conversational formulae, etc., among which idioms are considered the “prototype” of MWUs (Glaser 1998: 126). Influenced by the Russian tradition of phraseology, Cowie (1981) establishes a continuum, which goes from free combinations to pure idioms through restricted collocations and figurative idioms. Howarth’s model of MWUs (1998) has also been developed drawing on the work of Cowie (1998) and Aisenstadt (1981).

Despite the contribution to the field of phraseology, Russian phraseological theory has been criticized by other researchers. For instance, Pawley and Syder (1983) argue that most MWUs are not true idioms but rather regular form-meaning pairings. For this reason, they put forward a definition of *lexicalized sentence stems* as follows.

[A *lexicalized sentence stem* is] a unit of clause length or longer whose grammatical form and lexical content is wholly or largely fixed; its fixed elements form a standard label for a culturally recognized concept, a term in the language (Pawley and Syder 1983: 192).

As indicated from the definition, the lexicalized sentence stem is a culturally standardized designation (term) for a socially recognized conceptual category and it carries the authority of regular and accepted use by members of the speech community.

#### **2.1.1.2. Psycholinguistic approach**

Phraseology has also been touched upon the field of psycholinguistics. Evidence can be found from the study conducted by Wray (2002), where the terminology *formulaic language* is proposed:

[*Formulaic language* is] a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar (Wray 2002: 9).

Clearly, the term ‘formulaic sequence’ tends to cover various types of word combinations that vary in complexity and internal stability but appear to be stored and retrieved whole from memory at the time of use. Thus formulaic language “is or appears to be” prefabricated, i.e., being handled effectively like single “big words” (Ellis 1996: 111). Wray’s definition is frequently used for psycholinguistic investigations of phraseology. She and others (Bolinger 1976; Nattinger 1988) restrict the scope of phraseology to the ready-made memorized sequences that are typical of complete structures and idiomatic meanings.

### 2.1.1.3. Cognitive grammar

Unlike the above two approaches proposing a theoretical notion for understanding phraseology, Cognitive Grammar does not have a theoretical notion that is precisely equivalent of MWUs. Rather, it has a more general term, of which MWUs constitute a subset. By doing away with a strict separation between lexicon and grammar, Cognitive Grammar only contain *symbolic unit* in linguistic system, as defined below:

[A *symbolic unit* is] a structure that speaker has mastered quite thoroughly, to the extent that he can employ it in largely automatic fashion, without having to focus his attention specifically on its individual parts for their arrangement [...] He has no need to reflect on how to put it together (Langacker 1987: 57).

As can be seen from this definition, a symbolic unit is a pairing of form and meaning or function. If a language user encounters a particular symbolic unit quite frequently, this symbolic unit is likely to become be accessed automatically and entrenched in his/her linguistic system. Even though a symbolic unit is identical to that of an MWU, it is apparent that MWUs do not enjoy a special status within Cognitive Grammar, but only one subtype of symbolic unit.

In Cognitive Grammar, Fraser (1976) also defines *idiomatic expression*, which is similar to *symbolic unit*.

[An *idiomatic expression* is a] single constituent or series of constituents, whose semantic interpretation is independent of the formatives which compose it (Fraser 1976: v).

What different from *symbolic unit* is that Fraser's idiomatic expression is proposed from a discourse angle, and more importantly, it discusses the non-compositional feature of MWUs. In this sense, the notion of *idiomatic expression* implies the nature of phraseology, although recent studies have shown that non-compositional semantics

is not a necessary condition for defining MWUs (Goldberg 2006; Partington 2004; Svensson 2008). Instead, the frequency of a sequence, which is large enough for it to become entrenched, helps attain phraseological status (Goldberg 2006).

#### 2.1.1.4. Corpus-driven approach

A more recent approach to defining phraseology is the corpus-driven approach. The focus of this approach is typically on co-occurrences of word forms that are recurrent in authentic texts, drawing on Firth's concept of "meaning by collocation" (Firth 1957: 39).

Being assisted by advances in computer technology and the development of very large corpora, corpus-driven approach solves some complex problems of traditional definitions of MWUs. That is, instead of listing linguistic criteria, it uses a bottom-up approach to identify MWUs. And the significance of an MWU is calculated by means of statistical measurement, which the occurrences are more frequent than probability by chance (Sinclair 1991).

Using the corpus-driven approach, Grice (2008) put forwards a more rigorous definition of the *phraseologism* as follows:

The co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance (Grice 2008: 4).

In this definition, "frequency of co-occurrence" is considered as a principal criterion for defining MWUs. Based on this, the phraseologism is thus characterized by a sufficiently high frequency of co-occurrence, even when a strict threshold value is not provided. apart from that, Grice (2008: 3) further argues that "all linguistic inferences in the field of phraseology are dependent on statistical information of some kind." Specifically, the phraseologism includes the co-occurrence of a form of a



lexical item plus any other kind of linguistic element, which can be another form of a lexical item or a grammatical pattern. Clearly, the corpus-driven approach adopts a much wider perspective and encompasses many MWUs that would traditionally be considered to fall outside the scope of phraseology.

### **2.1.2. Parameters for defining MWUs**

It seems clear from the above survey of the terminologies on phraseology that there is not a clear-cut definition provided in this field. Thus, Gries (2008) points out the importance of explicating the defining parameters of MWUs in phraseological research, which allows other researchers to recognize more easily the potential areas of overlap, or conflict for that matter. By closer comparative look at the vast majorities of studies that exist, Gries also identified a set of parameters that typically underlies in phraseological research for defining MWUs. Following Gries (2008), the task for defining MWUs in this study is guided by these establishing parameters. In general, we hold that a rigorous definition of co-occurrence phenomena in general and phraseology in particular, needs to involve at least four parameters. They are co-occurrence of linguistic elements, length of co-occurrence, significance of co-occurrence in statistical measurement, the degree of inflexibility of co-occurrence, Structural cohesion of co-occurrence and semantic unity of co-occurrence.

#### **1) Co-occurrence of linguistic elements**

As MWUs are the co-occurrences of lexical items on the syntagmatic level, most phraseological studies concern with continuous sequences rather than the discontinuous one. (Altenberg 1998; Biber *et al.*, 1999; De Cock 1998, 2000; Eeg-Olofsson and Altenberg 1996; Jurafsky and Martin 2000; Manning and Schutze 2000; Stubbs and Barth 2003) In the present study, we adopted the same perspective and considered the co-occurrence between adjacent words to be a potential MWU.



## 2) Length of co-occurrence

Apparently, multi-word sequences entails that MWUs are made up of at least two words. Therefore, 'length' is widely regarded as one of the basic or primary parameters of defining MWUs (Altenberg 1998; Gross 1996; Knappe, 2004; Mejri 2005). Among phraseological studies, although some researchers have specified the length of MWUs that they want to include in the scope of their investigation (Altenberg 1998; Biber and Butler 1997; Biber *et al.* 1999; Conrad and Cortes 2003; Sugiura 2002), they only target at the MWUs consisting of three to five words, with 2-word sequences left out. This treatment, as Altengberg (1998: 103) admits, inevitably "excludes a number of phraseologically interesting idioms and collocations," since 2-word sequences have been proven to account for over 60% of the corpus (Piao, Rayson, Archer, Wilson and McEnery 2003: 54; Sinclair 2001: 353). For this reason, this study considers the issue of length as a defining dimension. However, for the convenience of automatic extraction from corpora, the length is only restricted to the continuous MWUs consisting of two to five words ( $n$ -grams, with  $2 \leq n \leq 5$ ). Any discontinuous MWU and the MWUs with more than five words are beyond the scope of the present study.

## 3) Significance of co-occurrence in statistical measurement

In corpus-based phraseological studies, a threshold of absolute frequency of co-occurrence is commonly used for defining MWUs (Altenberg 1998; Biber *et al.* 1999; De Cock 1998, 2000; Simpson 2004). One of the most representative works conducted by this approach is Biber's study of lexical bundles, in which the researchers use a cut-off frequency of 40 times per million words to extract the most frequent sequences of words in two university registers (Biber *et al.* 2004). Despite the fact that such method can reflect the "probabilities in the linguistic system" (Halliday 1993: 3) and detect what is typical in language (Stubbs 2002: 227), it has been convincingly shown that this approach has resulted in inevitably a lot of "phraseologically irrelevant examples" (Altenberg and Eeg-Oloffsson 1990: 16)

inflating the number of what could be considered phraseological in a corpus. Therefore it cannot be used as the sole criterion for MWU identification. To tackle this problem, some researchers suggest that the parameters used for defining MWU should take into account other issues. As Sinclair points out, MWUs should be “a string of lexical items co-occurring with mutual expectancy greater than chance” (Clear 1993; Sinclair 1991). This claim indicates that the association strength between the co-occurred words takes a more important role in MWU extraction. Other researchers also hold the same view by arguing that it is “probably better to use as much information as possible in exploring associations, and to take advantage of the different perspectives provided by the use of more than one measure” (Barnbrook 1996: 101).

Based on the above discussion, this research adopts a refined association measure as a primary statistic to ensure the identification of relevant co-occurrence data. That is, a sequence of words cannot be accepted as an MWU unless its observed frequency of occurrences is significantly higher than its expected frequency and the association within a sequence is statistically significant.

#### 4) The degree of inflexibility of co-occurrence

In the literature, MWUs have long been referred to as “fixed expressions”. The syntactic inflexibility is thus regarded as a determining factor of phraseological status and more particularly of idiom status. However, recent corpus-based research has shown that MWUs are in fact not all “fixed” (Sinclair 2004: 30) and the idea of the fixedness of form is false (Moon 1998: 47). Nowadays, researchers tend to treat the inflexibility of a sequence as a continuum and believe that fixedness is not a clear-cut dichotomy but a matter of more or less (Howarth 1998: 169; Langacker 1987; Svensson 2008; Van Lancker 2004: 4). The present study therefore followed this tolerant view to syntactic inflexibility considering it as an indication rather than a criterion for defining MWUs.

## 5) Structural cohesion of co-occurrence

One requirement presented in the literature of MWU identification is the structural cohesion within the constituents. This idea was put forward based on the fact that a syntactic relation between words in a word combination indicates a potential semantic relation. This relation can be assumed to become relatively stable and established in the language when the word combination occurs frequently in the same constellation and meaning. Kjellmer (1991: 116) is the first to emphasize this issue in his definition of collocations as “recurring sequences that have grammatical structure.” Simpson (2004) also holds similar view but set a strict structural requirement for MWU identification. That is, in order for a sequence of words to merit the status of being formulaic, they must “constitute complete syntactic units” (Simpson 2004: 42) or must “be relatively well-structured” (ibid.). In this study, we didn’t take Simpson’s approach to defining MWU. But we were approved of the idea that the constituents of an MWU should be directly and syntactically related. In another words, with this criterion, sequences that are highly recurrent but composed of weakly related syntactic units, are excluded, such as *and me when I*.

## 6) Semantic unity of co-occurrence

The semantic status of MWUs is one of the most important criteria for MWU identification. As regards to this issue, many studies have been devoted to understanding the semantic properties of the single words within the sequence. It has usually been achieved either by discovering their contribution to the overall meaning of the sequence, or by comparing the meaning and use of single words within and outside the sequence.

Among all the discussions, non-compositionality has often been mentioned. A multi-word sequence is said to be non-compositional if its meaning is different from the sum of its individual parts (Svensson 2008). This feature undoubtedly characterizes certain types of MWUs, such as idioms and locutions (Erman and Warren 2000; Gonzales-Rey 2002; Gross 1996; Hudson 1998; Moon 1998; Svensson

2004). However, it is not a necessary parameter for defining MWUs as a whole, since corpus data provide ample evidence to show that most MWUs are compositional by their nature (Stubbs 2001). Therefore in the present study, for a sequence of words to be counted as an MWU, semantic unity is required, but not necessarily non-compositional. In another word, non-compositionality is only considered to be an indication of any specific type of an MWU rather than a defining criterion.

Besides the criteria discussed above, the issue of pragmatic function is noteworthy. With divergent research aims, researchers give different treatments to this criterion. In Nattinger and De Carrico's (1992) research, only word sequences with pragmatic functions can qualify as a lexical phrase whereas in other studies, especially in the investigations of collocations, no such criterion is required. In the present study, this attribute is considered to be an optional criterion, that is, an FS may or may not be pragmatically loaded.

In brief, all the parameters proposed here underscore the linguistic dimensions that are relevant to phraseology and provide a framework of reference for defining MWUs. Surely, some researchers may propose different parameter settings depending on their research goals. Therefore they exclude some of these or include additional ones. But in the present study, these parameters guide the entire process of MWU identification and ensure the scope of the investigation operational. The working definition of MWUs will be brought forward in the next subsection.

### **2.1.3. Operational definition of MWUs**

With the foregoing in mind, this study defines an MWU as:

*A structurally relevant and semantically coherent multi-word unit, whose empirical internal association is significantly greater than expected on the basis of chance.*

This operational definition differs from other definitions in two aspects. First, it

requires the word sequence constitutes syntactically related structure and coherent semantic meaning. The requirement concerning ‘relevant syntactic relation’ was proposed on the basis of the previous literature on the MWU definition. For instance, the issue of “recurring sequences that have grammatical structure” was first put forward by Kjellmer (1991: 116) in his definition of collocation. Simpsons (2004) also discusses the criteria for being MWUs in a more explicit manner. That is, an MWU must be “relatively well-structured” or must “constitute complete syntactic units”, which include prepositional phrases (*at the end, in the past*), noun phrases (*a lot of people, the first thing, something like that*), verb phrases (*to make sure, look at this*), or entire clauses (*I can’t remember, does that make sense*) (Simpsons 2004: 42). The advantage of syntactic criterion, according to these researchers, is rooted in the fact that a syntactic relation between words in a word combination indicates a potential semantic relation. This relation can be assumed to become relatively stable and established in the language when the word combination occurs frequently in the same constellation and meaning. While the decision about whether or not an MWU is a semantically meaningful unit has to be made based on human judgment as methodological support (Simpsons 2004). This is because some word sequences cannot be viewed as semantic units in their own right, even though they are complete in form.

Based on these requirements, any word sequences, which are composed of weakly related syntactic units or do not intuitively look, sound and feel like semantically independent expressions, were considered noise and excluded from the set. This can be exemplified by word sequences such as *the accident the* (532), *accident the* (50), *figure however* (45), *for the vessel to* (43), etc.

Second, the operational definition puts an additional statistical emphasis to the MWU. To be more specific, in order to be qualified as an MWU candidate in this study, a multi-word sequence should be composed of two to five words and its internal association must be statistically significant.

It is obvious from the above proposed definition that not all phraseological sequences are in the scope of the current research. First, the definition excludes part of

lexical bundles investigated by Biber *et al.* (1999) or recurrent word combinations by Altenberg (1998), which are structurally irrelevant and semantically incoherent units, even though these sequences occur frequently in the corpus. Examples include *the accident the* (532), *accident the* (50), *figure however* (45), etc. Second, the requirements of this definition does not consider whether the multiword sequences are fixed or semi-fixed in structure, compositional or non-compositional in meaning and pragmatic or non-pragmatic in function. In another word, these criteria are optional in MWU identification in the current research.

The next section will focus on the theoretical account of phraseology, through which the linguistic nature of phraseology will be uncovered. Then a brief review of the previous studies under the rubrics of theory-driven research paradigm and corpus-driven paradigm will be provided at last.

#### **2.1.4. An overview of influential taxonomy of phraseology**

Similar to the issue of definition, there is a lack of a coherent classification framework for MWU analysis. Such incoherence has also created challenges for the studies of MWUs. For better understanding of this issue, this subsection reviews some of the most influential taxonomies developed to date, which is also believed to provide a theoretical foundation for the present study.

Over the past two decades, a variety of schemes of classification have been proposed with varying focuses (Aijmer 1996; Biber *et al.* 1999; Coulmas 1994; Cowie 1988; Erman and Warren 2000; Howarth 1998; Krashen and Scarcella 1978; Moon 1998; Nattinger and De Carrico 1992) and a number of distinct categories of MWUs have been identified. Obviously, a review of all these taxonomies is beyond the scope of this dissertation. So in this section we only review the typical ones, which form the basis of a proposed taxonomy for the present study.

A survey of the research literature on the classification of MWUs reveals that most researchers classified MWUs on both structural and functional grounds (Altenberg 1998; Biber *et al.* 1999, 2003; Cowie 1988; Erman and Warren 2000; Nattinger and



De Carrico 1992). For example, in Nattinger and De Carrico's (1992) taxonomy, four structural categories and three functional categories are identified. The four structural categories include polywords (e.g., *you see, so far so good*); institutionalized expressions (e.g., *how do you do*); phrasal constraints (e.g., *a \_ ago; dear \_*) and sentence builders (e.g., *I think \_; That reminds me of \_*). According to Nattinger and De Carrico (1992: 37-38), these four structural categories are different in four aspects: (1) length and grammatical status; (2) canonical or non-canonical; (3) fixed or semi-fixed; (4) continuous or discontinuous. The three functional categories are social interactions, necessary topics, and discourse devices. Social interaction markers consist of two categories: conversational maintenance such as summoners (e.g., *pardon me, hello, what's up ...*) and conversational purpose, such as offering (e.g., *Would you like \_?*). Necessary topic markers are lexical phrases which mark topics often discussed in daily conversation, such as "*My name is \_*", "*I'm from \_*". Discourse devices are those which connect the meaning and structure of the discourse, such as "*as a result of \_*", "*because of*" (Nattinger and De Carrico 1992: 60-65).

Nattinger and De Carrico (1992: 46) themselves draw attention to the limitations of this taxonomy, recognizing that their structural categories have fuzzy edges. For example, prototypical polywords (e.g., *at any rate*) will be completely invariable, whereas phrasal constraints will usually allow some variations. Between these poles, though, lies a fluid borderline, as evidenced by such polywords as *for better or worse*, which allows syntagmatic variation (*for better or for worse*). Furthermore, Nattinger and De Carrico (1992) treat only word sequences with pragmatic functions as members of lexical phrases. This treatment excludes the semantically literal and grammatically regular strings, such as collocations, from the scope of investigation. But a large body of corpus studies (e.g., Altenberg 1998) has evidenced that word sequences of this kind are pervasive in language use and constitute an important part of language users' competence.

In line with Nattinger and De Carrico (1992), Altenberg (1998) and Biber *et al.* (1999, 2003) also set up their own classification criteria for MWU analysis.

In Altenberg's taxonomy (1998), three broad categories of MWUs are distinguished

according to the grammatical structures. They are full clauses, clause constituents and incomplete phrases. Depending on the degree of completeness of grammatical structures, full clauses are further divided into dependent clauses and independent clauses while clause constituents are further divided into multiple clause constituents and single clause constituents. When dealing with functional classification, Altenberg discusses it under structural categories, which exhibits differences from other researchers. To be specific, each structural sub-category of MWUs is further classified according to the functions they perform in discourse. For example, independent clauses are classified into categories of responses, epistemic tags and meta-questions; from a textual perspective, multiple clause constituents are divided into seven different functional categories depending on their function and position in the linear organization of the clause: frame, onset, stem, medial, rheme, tail and transition. It is clear that the functional classification of this kind is, to a large extent, dependent on individual investigator's intuition and the decision is consequently difficult to keep consistent between different researchers.

Compared to Altenberg's (1998) taxonomy, Biber *et al.*'s (1999, 2003) classification scheme for lexical bundles is more detailed and easy to operate. Although most lexical bundles do not represent complete structural units, yet they show strong grammatical correlates. Hence, Biber *et al.* (1999: 996) group them into 14 major structural categories. On the basis of the most frequent bundles identified in Longman Spoken and Written English Corpus (Biber *et al.* 1999), Biber *et al.* (2003) and Cortes (2002) propose a functional classification scheme. The four core categories in their taxonomy are: referential bundles, text organizers, stance bundles and interactional bundles, each of which has several sub-categories associated with more specific functions and meanings. It is obvious that these four categories are closely related to the linguistic functions described by Halliday (1994). Referential bundles perform an ideational function. They make direct reference to elements in the physical world. Text organizers are word combinations used to express textual functions. They reflect relationships between prior and coming discourse. Stance and interactional bundles perform interpersonal functions. Stance bundles express



attitudes or assessments of certainty towards the following proposition. Interactional bundles are conversational word combinations used to express politeness or to report.

Clearly, the discussion of MWUs from lexical bundle perspective brings some limitations, one of which is it does not take the discontinuous word sequences into consideration, thus cannot serve the purpose of investigating discontinuous MWUs. Another problem of Biber's taxonomy is hidden in the extraction method of lexical bundles. As Sinclair points out, the crude method of retrieving lexical bundles inevitably leads to the inclusion of some meaningless sound bundles (Sinclair 2001: 353; 2004b: 10). For analyzing the MWUs which are structurally coherent and semantically complete, the classification scheme needs to be refined.

Erman and Warren (2000) classify all MWUs, prefabs in their terminology, into four large categories: lexical, grammatical, pragmatic, and reducible. In their taxonomy, lexical prefabs and pragmatic prefabs are further classified into different subcategories according to the syntactic and functional features, respectively. According to Erman and Warren (2000), lexical prefabs are semantic units, which denote entities, properties, states, events, etc. Examples are "*rules of sth.*", "*maths and physics*", "*a waste of time*", "*the present state of our knowledge*". Lexical prefabs are further divided into phrase type and clause type according to their structural properties. Pragmatic prefabs refer to word sequences that do not directly partake in the propositional content of the utterance in question. They are highly conventionalized and typically correspond to specific interactional situations, such as "*I'll talk to you later!*" (corresponding to the end of an exchange), and "*Enjoy your meal!*" (describing what precedes the process of eating between guests). Grammatical prefabs are intra-linguistic text-forming items rather than units with extra-linguistic reference, such as "*a little bit*" as a quantifier, "*be going to*" for tense-forming, and "*might be*" for mood-forming.

The advantage of Erman and Warren's scheme is that it broadens the scope of MWU research, which is evident in two aspects. Firstly, it includes the MWUs which are not pragmatically determined. Secondly, it introduces the word sequences with meta-linguistic functions into the area of MWU research, such as "*and everything*",

“sort of”, etc. This treatment provides us with more chances to have a better understanding of MWU phenomena. However, this scheme is not without problems, especially in the determination of whether a prefab is lexical or grammatical and sometimes whether it is a grammatical or pragmatic one. In addition, the structural analysis is only applied to lexical prefabs, leaving pragmatic prefabs not described formally.

In general, the existing taxonomies of phraseology provide general frameworks for MWU analysis from different dimensions. For the purpose of the present study, we proposed a refined taxonomy by synthesizing the taxonomies reviewed above, as explicitly discussed in Chapter 3.

## **2.2. Theoretical discussion of MWUs**

### **2.2.1. Theoretical framework of this study**

This study is theoretically set in Firth’s linguistics, in which the Firth’s ‘contextual theory of meaning’ is one of the central assumptions. According to Firth,

*The complete meaning of a word is always contextual, and no statement of meaning apart from a complete context can be taken seriously (Firth 1957: 7).*

Clearly, this argument indicates that the meaning of a word depends on its context. Based on this, meaning can be perceived as a complex of contextual situational relations. It links all linguistic elements (from phonetics to lexicography) with their context and situations. Understanding ‘meaning’ in this way allowed researchers to explore ‘meaning’ at each linguistic level (Chapman and Routledge 2005). For instance, meaning embraces the notion of the ‘collocation’ at the lexical level and the notion of colligation at the grammatical level:

*Collocation is the occurrence of two or more words within a short space of each other in a text (Sinclair 1991: 170).*

*The statement of meaning at the grammatical level is in terms of word and sentence classes or of similar categories and of the inter-relations of those categories in colligation... (Firth 1957: 13).*

It is this meaning-oriented approach that guides the linguists to discover the fact that “most of words have no meaning in isolation, or at least are very ambiguous, but have meaning when they occur in a particular phraseology” (Hunston and Francis 2000: 270).

Another theoretical background of the present study is that a corpus enables us to discover the interconnections between grammar and vocabulary. Just like Lindquist (2009: 51) puts it, ‘lexical items or small classes of lexical items not only have their own meaning but also their own “local” grammars’. A review of relevant literature shows that a growing number of studies provides ample evidence for the inseparability of lexis and grammar (Francis 1995; Gries 2008; Hoey 2005; Hunston 2002, Hunston and Francis 2000; Partington 1998, Romer 2005, 2009; Scott and Tribble 2006; Sinclair 1991, 2004; Stubbs 2001; Tognini-Bonelli 2001) and the contributions to Granger and Meunier (2008), Meunier and Granger (2008), Romer and Schulze (2008, 2009). Among them, one typical example of such integration is the ‘collocational frameworks’ proposed by Renouf and Sinclair (1991), which are ‘composed typically of a sequence of small closed word classes and/or individually specified members of such classes’ (Sinclair 2008: 408). In addition, Francis (1995) also argues that particular syntactic structures tend to co-occur with particular lexical items and — the other side of the coin — lexical items seem to occur in a particular range of structures.

Under the above theoretical backgrounds, phraseological studies are believed significant for gaining insight into the nature of language.

### **2.2.2. Nature of multiword units**

As we discussed above, phraseology has been of interest in many areas of

linguistics, such as Pattern Grammar and Corpus Linguistics. Although they explore the phenomenon from different perspectives, with diverse methods and under various labels for reference, they are consistent in understanding the nature of MWUs as form-meaning pairings, in which “meaning” represents both semantic content as well as pragmatic functions (Croft 2001: 19).

In this section, we will explore the meaning aspects of MWUs from the perspective of Corpus Linguistics. Within this field, MWUs are semantically defined as extended units of meaning by Sinclair (1991, 2004a) and pragmatically defined as by Nattinger and De Carrico (1991).

#### **2.2.2.1. MWUs as extended units of meaning**

Traditionally, individual words have been regarded as the primary units of meaning, which can be confirmed by a glance at the entries at any dictionary. However, findings from recent corpus-based studies have indicated that this is not the case. Instead, it has been proposed that units of meaning are “largely phrasal” (Sinclair 2004: 30). As Sinclair (2008: 409) notes, “however we circumscribe the unit of meaning, there will be connections like tentacles stretching out to the surrounding context, supporting or modifying the selection. We have to concede that the normal primary carrier of meaning is the phrase and not the word. The word is the limiting case of the phrase, and has no other status in the description of meaning”.

This claim can be supported by several evidences, such as the phenomenon of polysemy and what Sinclair (1991: 113) calls “a progressive de-lexicalization.” For polysemous words, even though the senses are distinguished by dictionary, the word alone is ambiguous or indeterminate in meaning. Therefore, it does not constitute a unit of meaning. Ambiguity can be eliminated only when the word occurs either in a physical context, or in a co-text, with other words around it. This phenomenon can be exemplified by the word *bank* in Stubbs’ (2001) study, in which he investigated all occurrences of *bank* (n=82) and *banks* (n=28) in the Lancaster-Oslo/Bergen Corpus of British English (LOB), an one-million written English corpus. By looking at the

concordance lines of the word, the researcher found that the word occurs mostly in fixed phrases which signal unambiguously the “money” or “ground” sense. In addition, the word usually co-occurs with other words which clearly signal one or other semantic field, which leads to greater regularity of collocation. In general, the ambiguity of meaning is reduced in linguistic contexts.

In everyday English, the phenomenon of de-lexicalization is much more common than expected. For example, Stubbs’ (2001) investigation of ‘*take a*’ in a corpus of over two million words shows that this lemma pair is commonly used in combinations such as *take a close look at*, *took an interest in*, *take a deep breath*, *take a photograph* and *take a decision*, where *take* is de-lexicalized. In fact, the corpus-based evidence illustrates that only 10 percent of over 400 examples does *take* have a literal meaning of “grasp with the hand” or “transport”.

The above-mentioned studies support the belief that individual words are not independent units of meaning. Rather, combinations of words in phrases seem to be a good candidate for the basic semantic unit of language in use. The systematicity of the co-selection of a word and its environment has led Sinclair to propose the notion of “extended unit of meaning” (Sinclair 1991: 24) or “lexical item” (Sinclair 1991: 141-148), for better presenting the primary unit of the meaning. for further clarification, Sinclair (2004) also puts forward five categories of co-selection as components of an extended unit of meaning: *the node word, collocation, colligation, semantic preference and semantic prosody*. As Sinclair (1998: 142) points out, the node word is “invariable, and constitutes the evidence of the occurrence of lexical item as a whole” (Sinclair 1998: 141); while the other four categories describe four types of meaningful relations pertaining to an extended unit of meaning, which are explained by Stubbs (2009: 22) in the following way (the wording has been slightly simplified):

COLLOCATION *is the relation between a word and individual wordforms which co-occur frequently with it.*

COLLIGATION *is the relation between a word and grammatical categories*

*which co-occur frequently with it.*

SEMANTIC PREFERENCE *is the relation between a word and semantically related words in a lexical field.*

SEMANTIC PROSODY *is the discourse function of the word: it describes the speaker's communicative purpose.*

In brief, the extended unit of meaning is a kind of unit based on a lexical core and extended to incorporate grammatical as well as other lexical choices. Through their lexical (collocation), syntactic (colligation) and semantic (semantic preference) flexibility, the units allow for a limited paradigmatic choice and thus an integration with other extended units of meaning in their context. New meanings are created when contextual constraints and lexical specifications do not match.

Sinclair hints at the fact that there may be a second kind, although it needs further study. He discusses this second kind of unit of meaning under the name of “collocational frameworks” (Renouf and Sinclair 1991), and argues that it is based on a grammatical core rather than a lexical one, usually discontinuous, such as *the...of*. Sinclair's idea on the second kind is taken up and further developed into the concept of (grammatical) pattern in Pattern Grammar by Hunston and Francis (Hunston and Francis 1998, 2000), which will be discussed in the next section.

#### **2.2.2.2. Pattern and meaning**

The notion of ‘pattern’ has been used as the meeting point between lexis and grammar. An example of this approach is given by Hunston and Francis (2000), who define their Pattern Grammar as a corpus-driven approach to the lexical grammar of English. The definition of what counts as ‘a pattern’ is reported below.

*The patterns of a word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently,*



*if it is dependent on a particular word choice, and if there is a clear meaning associated with it* (Hunston and Francis 2000: 37).

It is this notion that indicates pattern and meaning are co-selected: each pattern tends to be associated with certain meanings, realized by a restricted set of lexical items, and each lexical item tends to occur within a restricted set of patterns (Hunston and Francis 2000: 3). For instance, the function of the pattern ‘it v-link ADJ of N to-inf’ (i.e., *it was nice of you to come, it was courageous of him to speak out*, etc.) is to evaluate the action indicated by the *to*-infinitive clause, since the adjectives used in the pattern indicate judgment of an action (Francis, Hunston and Manning 1998: 501-502).

The same issue has also been discussed by Goldberg (1995) within the framework of Construction Grammar. She argues that constructions, similarly defined as “form-meaning correspondences themselves carry meaning, independent of the words in the sentence” (Goldberg 1995: 1).

In brief, this section lists the work which provides substantial and well documented evidence about form-meaning relations. In this section we have reviewed the linguistic treatments of MWUs in the corpus-driven description of English, specifically in the notions of extended unit of meaning and grammatical pattern. Literature reviewed has well documented that both notions are premised on the belief that syntax is not distinguished from lexis, hence form and meaning are associated. In the present study, we argue that both notions are linguistic realizations of form-meaning pairings, but they reflect two different ways of identifying and defining meaning. Extended unit of meaning reflects the way that begins with a node as a core, and then extends to integrate the contextual and functional information into one unit while grammatical pattern reflects the way that begins with a statement about the grammar and goes on to refine this to provide a description of a particular word. Both ways of analysis lead to the identification of lexico-grammatical patterning. These two ways are consistent with what Stubbs (2005) calls two strategies for meaning analysis: from lexis to co-text and from n-grams to content. In the present study, we

followed both ways to investigate the semantic features of MWUs in the genre of MAIR.

### **2.2.2.3. MWUs as form-function composites**

Apart from entailing the semantic meanings, MWUs also serve pragmatic functions. This nature is clearly reflected in Nattinger and De Carrico's (1992) definition of lexical phrases as form-function composites.

From the point of view of Nattinger and Decarrico, FSs, i.e., lexical phrases in their terms, are more than specific strings, they are also assigned functional meanings, so that these strings "not only have syntactic shapes, but are capable as well of performing pragmatic acts" (Nattinger and Decarrico 1992: 11), such as promising, complimenting, asserting, and so on. From authentic corpus data, Nattinger and De Carrico identify three functions instantiated by three large categories of lexical phrases: social interactions, necessary topics, and discourse devices. So, lexical phrases play particular functions in particular social contexts, and certainly constitute part of communicative competence.

In this section, we have mainly dealt with the nature of MWUs as form-meaning pairings. Ample evidence from corpus analysis confirms that all MWUs possess not only syntactic and morphological forms but also specific meanings. Further, the specific meanings entail not only traditional semantic interpretations but also pragmatic functions. Hence, MWUs represent the interface of linguistic and pragmatic competence and embody linguistic form, meaning and function.

The discussion on the nature of MWUs has suggested that, besides the amount of use, an adequate description of MWUs should also cover form, meaning and function. This observation lays theoretical foundation for the establishment of a conceptual framework for the present research.

### **2.2.3. Previous studies of phraseology**

Beginning in the 1990s, most phraseological research has been empirical, utilizing



corpus analysis to investigate MWUs in natural language. In general, two major methodological approaches have been employed: corpus-based and corpus-driven, terms originally introduced by Tognini-Bonelli (2001: 84-87). Corpus-based studies typically use corpus data in order to explore a theory or hypothesis, typically one established in the current literature, in order to validate it, refute it or refine it (McEnery and Hardie 2012). The methodological steps that are usually taken in the studies of this type include that the researchers first pre-select the MWUs that are perceptually salient or theoretically interesting, and then analyze the corpus to discover how those expressions are used (Moon 1998). Clearly, the definition of corpus linguistics as a method underpins this approach to the use of corpus data in linguistics (Tognini-Bonelli 2001: 84-85). In contrast, corpus-driven research paradigm does not start from theoretical models of language. Rather, it is believed that corpus itself embodies theories of language. Based on this view, researchers treat corpus as the only source of their hypotheses about language. As reflected in phraseological studies, the MWUs that are noteworthy are identified solely from the inductive analysis of a corpus instead of human judgment. Although the extracted MWUs may not fit any predefined linguistic categories, it has opened up a “huge area of syntagmatic prospection” (Sinclair 2004:19).

In general, the corpus-based versus corpus-driven dichotomy creates a basic, binary distinction, under which most phraseological works can be sorted into one or the other group. While in collocational research, a hybrid approach that combining both two methods is usually employed instead, through which researchers begin with a theoretically interesting target word (or a set of roughly synonymous target words), and then explore the corpus to identify the collocates that frequently occur in the context of the target words. As the current research attempts to explore the phraseological features of the domain-specific MWUs in MAIR, a hybrid approach is therefore believed to achieve such purpose by identifying and describing the full set of the domain-specific MWUs that are prevalent in the corpus.

This section undertakes a survey of some of the most important corpus investigations of phraseology carried out to date, based on the above-mentioned

approaches. Particular attention will be paid to the relevant corpus-driven studies.

### **2.2.3.1. Corpus-based studies of phraseology**

A growing number of studies on the Michigan Corpus of Academic Spoken English (MICASE) relate in some way to the topic of formulaic language—namely, Swales and Malczewski (2001) on ‘New Episode Flags’ like *okay, so now*; Mauranen’s studies of metalanguage (2001) and formulae (2003), Poos and Simpson’s (2002) study of the hedges *kind of* and *sort of*, and Swales’ (2001) article on *point* and *thing*, in which he mentions phrases like *at this point* and *the thing is*. Other researchers have studies formulaic language in academic writing (Cortes 2002; Cortes *et al.* 2002; Oakey 2002), and have found that academic writing is rich in discourse structuring and stance expressions, some of which overlap with other spoken and written registers, and others of which seem particularly characteristic of academic prose.

### **2.2.3.2. Corpus-driven studies of phraseology**

While there have been relatively few corpus-based studies of MWUs, there have been numerous studies conducted based on corpus-driven approach. One of the earliest corpus-driven studies of MWUs was Salem’s (1987) analysis of repeated lexical phrases in a corpus of French government resolutions. In the late 1990s, corpus-driven studies of recurrent lexical phrases in English registers began to appear, which followed by a growing number of studies in the next few decades. Although these studies demonstrate some differences in terms of their overarching research goals, the role of register in the analysis and the nature of multi-word units, they are all contributed to the overall understanding of the phraseological features in natural language. This subsection reviews some of the most influential researchers in this field, together with their representative works.

#### **Sinclairian studies and views**

Sinclair is one of the strongest proponents of the view that phraseology is central to

our understanding of language, and not something belonging in the periphery (Ellis 2008).

These principles are further supported by a large number of studies. For example, Erman and Warren (2000) estimate that over half of fluent native text is constructed according to the idiom principle.

Phraseology is central throughout Sinclair's research: phraseological items, whatever their nature, take precedence over single words (Sinclair 2008: 408). Unlike researchers of the corpus-based approach to phraseology, Sinclair and his followers are much less preoccupied with distinguishing between different (sub)categories of MWUs or setting clear boundaries to phraseology.

### **Altenberg's Recurrent Word Combinations**

Altenberg (1998) was perhaps the first researcher to investigate frequently occurring lexical phrases in spoken English. By identifying 470 three-word sequences that occurred at least ten times in the London-Lund Corpus, he observes that the use of more or less prefabricated expressions exists at all levels of linguistic organization ranging from discourse level to smaller units acting as single words and phrases. Altenberg (1998: 121) also uncovers that, in his own word, "there are comparatively few examples that are completely frozen, semantically or grammatically. Rather, the great majority of the examples occupy a position along the cline between fully lexicalized units and free constructions." In other words, word combinations of this kind illustrate very clearly the difficulty (or impossibility) of making a sharp distinction between lexicon and grammar.

In addition to these two fundamental findings, Altenberg (1998) further claims that multiple clause elements tend to appear in recurrent clusters, reflecting the conventionalized ways of unfolding and presenting information in continuous discourse. These clusters can be seen as "interlocking building blocks" of differing size and meaning, and although their combinatorial possibilities are constrained by various factors, pragmatic, semantic, or grammatical, they represent an important

phraseological resource in speech production.

One of the great contributions that Altenberg makes to the field of phraseology is to tackle the complex structural characteristics displayed by recurrent word combinations. According to their grammatical characteristics, he distinguishes three broad categories of structures for recurrent word combinations falling into the categories of full clauses, clause constituents and incomplete phrases. Depending on the degree of structural completeness, full clauses are further divided into dependent clauses and independent clauses while clause constituents are subdivided into multiple clause constituents and single clause constituents. Along with the structural analysis of MWUs, the researcher also discussed some of the major discourse functions served by these expressions (e.g., as interactional responses, epistemic tags, and comment clauses), however, the functional classification is embedded into each of the broad structural categories.

As a seminal study of phraseology in the corpus-driven paradigm, Altenberg's research inevitably suffers from limitations, particularly in the extraction method of recurrent word combinations. He defines a "recurrent word combination" as any continuous string of words occurring more than once in identical form. Thus many of these sequences consist of mere repetitions or fragments of larger structures (e.g., *and the, out of the*), and hence are of little phraseological interest. Therefore "for practical reasons", Altenberg (1998: 102) limits his examination to word combinations consisting of at least three words occurring at least ten times in the corpus. He admits that these limitations are to a large extent arbitrary. The length restriction was chosen partly to reduce the number of fragmentary sequences, but mainly to reduce the material to a manageable size. However, it is argued that two-word sequences are the most common type of MWUs, accounting for more than half of the total number, and that by excluding them it also excludes a number of phraseologically interesting phrases and collocations (e.g., *part of, at least*). In addition, Altenberg (1998) lays his emphasis on a holistic description of structural categories from phraseological, grammatical, semantic, and pragmatic angles; whereas he fails to go into any further detail about the description of function. Even his categorization of functions depends

largely on individual researcher's intuition, and the decision is consequently difficult to keep consistent among different researchers.

### **Biber's Lexical Bundles**

Around the same time, Biber *et al.* in the 1999 Longman Grammar of Spoken and Written English discuss what they call lexical bundles in some detail, comparing academic prose to general conversation.

In their study, lists are provided for common four-word, five-word, and six-word lexical bundles, defined as sequences of words that occurred at least ten times per million words in the target register, distributed across at least five different texts. These bundles were also interpreted in structural/grammatical terms, just like Altenberg does. These structural correlates were significant in two respects: (i) most lexical bundles are not complete structural units. Rather, it is found that only 15% of lexical bundles present in conversation are recognizable as complete units, and (ii) most bundles bridge two different structural units. Another surprising finding was that almost none of these most frequent lexical phrases were idiomatic in meaning, although they could be interpreted as serving important discourse functions.

Afterwards, Biber (2006) studies "lexical bundles" in spoken and written university registers with the purpose of exploring the patterns of register variation. By identifying the phraseological features that are especially prevalent in particular registers, Biber's study is significant in providing evidence for the existence of systematic patterns of use across university registers and academic disciplines. Firstly, the study differentiates between spoken and written mode in terms of phrasal patterns. In general, the academic writing is characterized by the wide use of simple main clause and complex noun phrases and prepositional phrases. However, such syntactic structures are rare in speech. According to Biber, a reason behind the pattern differences is the situational context of use. In spoken registers, speakers reveal personal feelings and attitudes face to face, while in written registers, writers address a more general distanced audience of readers.

Secondly, as regards the discourse functions of lexical bundles, the evidence from Biber's research indicates that lexical bundles expressing stance meanings are frequent and pervasive in the university language. Moreover, compared with the use in spoken registers, stance lexical bundles are more common in the "course management," a subdivision of the written register.

Overall, same as Altenberg (1998), Biber sets up prescribed frequency threshold as the only criteria for lexical bundle identification. This crude method of retrieving lexical bundles inevitably leads to the inclusion of a large number of disturbing segments such as *to go ahead and*, *to look at the*, *and in the*, and the exclusion of a number of sequences that are of particular interest in phraseology, such as *in further research*, *in sharp contrast*, and *in a word*.

Numerous subsequent studies have employed a "lexical bundle" framework to describe the lexical expressions typical of different registers, focusing on both frequency and discourse function. The most recent work in that tradition has been done by Jhang, Kim and Qi (2018), in which the authors compare the construct of lexical bundles by L1-English versus L1-Japanese professionals in the genre of marine accident investigation reports (MAIR). It is found that compared with English reporters, Japanese professionals employ a considerably wider range of four-word lexical bundles, exhibit an overuse tendency in almost all structural patterns and functional types and adopt different strategies to construct lexical bundles and fulfill discourse functions.

Recurring sequences of words have long been considered as a signifier of different genres and registers by corpus linguists (Biber and Barbieri 2007; Biber *et al.* 2004; Chen and Baker 2010; Cortes 2004), since Biber *et al.* (1999) observed that the internal linguistic features of lexical n-grams are different in conversation and academic prose. Biber *et al.* (2004) analyzed the frequencies, structural types and functional categories of n-grams and their distributions in university teaching and textbooks, and was extended by Biber and Barbieri (2007) to a wider range of spoken and written university registers. Cortes (2004) made a comparison between publications and student writings in history and biology. Chen and Baker (2010) did



structural and functional analysis of n-grams in corpora of L1 and L2 academic writing. Besides, Gries (2010a, 2010b, 2011) explored the n-gram frequencies among various registers with several advanced quantitative methods. The previous research mainly focused on lexical n-grams. Nevertheless, n-grams of other linguistic features, such as part-of-speech, have been much less studied (except Santini 2004). Santini (2004) presented genre classification experiments using unigrams, bigrams and trigrams obtained from BNC, and trigrams gained the best performance.

To sum up, corpus-driven research paradigm has brought about many landmark studies of phraseology with different features (Altenberg 1998; Biber 2006; Hunston and Francis 2000, 2008; Sinclair 1991, 2004, 2008). However, studies of this type are mostly restricted to general English texts, whereas there are few systematic studies of phraseology in specialized corpus, especially in maritime domain.

This chapter has given a brief overview of the field of phraseology. First, the theoretical background of MWU were addressed, including the theoretical framework of the studies and the nature of MWUs. Then, the previous studies of MWUs were reviewed, grouped based on corpus-based and corpus-driven paradigm. This has paved the way for the current research. In the next chapter, we will introduce the analytical framework and the research design for the current research.



## Chapter 3

### Analytical Framework and Research Design

This chapter focuses on presenting the analytical framework and methodology of the current research. It first introduces the analytical framework of this study followed by the research questions to be addressed (Section 3.1 and Section 3.2). Then it turns to provide a detailed description of the corpora used for the present investigation as well as the tools and procedures for data analysis (Section 3.3 and Section 3.4).

#### 3.1. Analytical framework

The nature of MWUs as form-meaning pairings indicates that MWUs are assigned both formal properties and meaning. As discussed previously in Chapter 2, the term “meaning” in this concept not only refers to the conventional semantic knowledge, but also implies the functional meaning of MWUs that used in specific pragmatic contexts (Croft 2001: 19). Therefore, it is considered inadequate to characterize MWUs exclusively from one dimension. In other words, only by a systematic description of phraseological features from perspectives of form, meaning and functions can we have an overall understanding of the nature of MWUs. As Stubbs stated (2005), “a description of a phrasal construction must state its internal and external features and provide a structural analysis and a functional analysis (of its meaning and communicative purpose).”

Based on this, the present study proposed a three-dimensional analytical framework to understand the use of domain-specific MWUs in MAIR genre. That means investigation does not only include the syntactic and semantic interpretations of MWUs, but also the analysis of their pragmatic functions. At this point, it is worthwhile noting that this three-dimension framework is designed mainly for the convenience of discussions but not for separating MWUs into individual modules, since the current research takes a holistic approach to the investigation of MWUs. The

following part demonstrates in details how the framework was implemented in this investigation.

### 3.1.1 Analytical framework for syntactic features of domain-specific MWUs

As regards the syntactic features of MWUs, it has been suggested to be analyzed under certain framework. For example, Simpson (2004: 38) argues that MWU use, as a co-occurrence phenomenon in syntagmatic level, is inevitably constrained by syntactic relations. Therefore, it is better described and investigated in certain syntactic frameworks. In fact, a review of relevant literature shows that a number of studies have been devoted to establishing structural frameworks for MWU analysis, all of which can serve as a guide for MWU research (Altenberg 1998; Biber *et al.* 1999; De Carrico 1992; Erman and Warren 2000; Hunston and Francis 2000; Nattinger, Pawley and Syder 1983; Wray 2000, 2002).

In the present study, an initial attempt was made to apply the “lexical bundle” framework to describe the different grammatical correlates of MWUs. The choice of this classification as the basis is primarily due to the practical reason, as it is much more intricate and convenient to operate compared with other taxonomies. However, by classifying the target MWUs, it was found that a set of word sequences could not fit into the categories, such as sentence stem patterns, *etc.* This is probably a result of different methods for extracting MWUs, and the specialized nature of the study corpus. Therefore, the present study synthesized the existing taxonomies for handling the extracted data from the Corpus of Marine Accident Investigation Reports (thereafter COMAIR).

In the final structural classification scheme, three major categories were included, namely, NP-based, VP-based and PP-based construction, following other researchers such as Chen and Baker (2010). NP-based and PP-based structures include noun phrases and prepositional phrases, while VP-based constructions refer to “word combinations with a verb component” (Chen and Baker 2010: 35). Then all three categories were further classified into several subcategories, as they usually present

complex structural patterns and variations. For instance, the categorization of NP-based structure was subdivided into two types of construction: 1) noun phrases and 2) NP-based fragments (i.e., NP + prepositions), since the target MWUs contain a large proportion of this structure as well. Similarly, the category of PP-based structure was also further classified into construction starting with *of* and construction starting with other prepositions. By contrast, VP-based structure displays more structural variation. To be specific, a brief examination of the MWUs of this type showed that although many of the word sequences can fall into the Biber *et al.*'s (1999) classification of VP-based structure (i.e., verb phrase with active verb; passive verb + (PP) fragment; copula *be* + PP fragment; *to*-clause fragment, *etc.*). There are a number of expressions serving as clause constituents in the COMAIR, such as sentence stems (*the master decided, regulations require*). To cover the clause fragments, the scheme was then supplemented by clause constituents proposed by Altenberg (1998). Hence, the VP-based category was broadly divided into VP-based fragments and clause fragments, within which several subcategories included. At last, some MWUs possess the structure of adjective phrases or adverbial phrase fragments. The MWUs of this type were assigned into the category of 'others'. Table 3.1 presents the structural taxonomy of the domain-specific MWUs specifically designed for the present study.

**Table 3.1 Structural taxonomy designed for present study**

Structural category	Subcategory	Examples
NP-based structure	a) noun phrases	<i>engine room; bridge team</i>
	b) NP-based fragments	<i>the course of; courses in</i>
VP-based structure	VP-based fragments	
	a) verb phrase with active verb	<i>left the bridge; informed the master</i>
	b) passive verb +(PP) fragment	<i>was loaded, been secured</i>
	c) copula <i>be</i> + adj./noun./pp.	<i>had been on board; was on passage</i>
	Clause fragments	

	d) sentence stems	<i>code requires; there was no requirement for</i>
	e) <i>to</i> -clause fragment	<i>to alter course; to be fitted</i>
	f) <i>that</i> -clause fragment	<i>estimated that; should ensure that</i>
	g) <i>it</i> -clause fragment	<i>it is evident; it is probable that</i>
<b>PP-based structure</b>	a) PP starting with <i>of</i>	<i>of collision, of fishing vessels</i>
	b) PP starting with other prepositions	<i>in the engine room, in accordance with</i>
<b>Others</b>	adjective or adverbial phrase fragment	<i>ahead and astern; dead slow</i>

In order to ensure the reliability of the structural classification, two researchers undertook the task together. Cohen's Kappa was used, yielding a result of 0.89 (the raters were highly consistent). The in-depth structural analysis of the target MWUs will be discussed in the subsequent Chapter 5.

### 3.1.2. Analytical framework for semantic features of domain-specific MWUs

The second dimension of analysis concerns the semantic interpretation of the domain-specific MWUs. It is commonly regarded as the essential aspect of MWU research, as Sinclair (2004a) states in his notion of 'extended unit of meaning' every unit of meaning has its own structure, which implies that form and meaning are co-selected and interwoven together. To achieve this, the present study adopted a lexical-grammatical approach proposed by Stubbs (2005: 5).

According to Stubbs, the semantic analysis of MWUs can be carried out from two perspectives: from lexis to co-text and from n-grams to content. As the names imply, the first perspective starts from selected individual words and studies their typical co-text, while the second perspective starts from recurrent n-grams and studies their typical content (Stubbs 2005). Among the previous phraseological studies, the second strategy has been mostly applied by researchers. One of the typical examples is the study of lexical bundles, which investigates the semantic and pragmatic meaning of lexical bundles by extracting n-grams first. For the first strategy, although its application is relatively fewer, it still can be found in a variety of research, such as

studies of the overall phraseology of individual words; studies of phraseology around set of words (Mahlberg's (2005) study of high frequency 'general' nouns and Lindquist and Levin's (2005) study of nouns from semantic sets) and studies of grammatical constructions (Stubbs 2006). Analyzing semantic meaning of MWUs in this way, as illustrated by Stubbs (2005), is based on the hypothesis that frequent words are frequent because they occur in frequent phrasal constructions, which express essential semantic and pragmatic meanings (Stubbs 2005). In other words, phrasal construction, as the lexical realization of a word, provides a way of expressing meaning.

Considering the subject of the current research is the phraseological patterns which are specific to MAIR genre, the MWUs formed around keywords were chosen to represent these patterns. The choice of keywords as the node words was by no means arbitrary, but principally determined by their significant role in the corpus. As is known, keywords can be indicative of the aboutness and stylistic features of the text, it is thus believed that the MWUs with keywords can best reflect the special phraseological features of MAIR genre.

To investigate the semantic features of the target MWUs, it is appropriate to adopt the above two perspectives in the analysis. The individual words required in the first perspective are selected from the keywords (lexis) whose semantic meaning displays difference from their use in general English register. Through investigation of the patterns of meanings realized by MWUs around keywords (co-text), the semantic features of the domain-specific MWUs in MAIR can be best understood. Similar to the first perspective, the second perspective for semantic analysis also allowed us to target at the word sequences (n-grams) presenting distinctiveness in the MAIR genre. By examining the concordances of these MWUs (content) and comparing with their use in general English register, their way how their use diverges from general English was clearly demonstrated.

### 3.1.3. Analytical framework for functional features of domain-specific MWUs

Being another meaning aspect, the functional features of domain-specific MWUs also deserve intensive investigation. Similar to the syntactic analysis, it is necessary to establish a functional taxonomy, into which these extracted MWUs could be divided. As reviewed in chapter two, there exists a variety of classification schemes offering promising methods for functional analysis of MWUs. Most of these frameworks are developed to characterize the functional use of MWUs either in general English or in EAP context. For example, in the framework devised by Biber *et al.* (2004) and Biber (2006), the subject-specific MWUs are excluded for the reason that they just reflect the immediate concerns of a particular text rather than a range of disciplines (Biber 2006: 175). Such a treatment obviously contrasts with most ESP studies, where the subject-specific MWUs are the main research interest. Although a number of researchers in ESP domain attempted to modify the established framework by adding a category to present the topic-specific functions (e.g., Jhang and Lee 2013b; Jhang *et al.* 2018), these studies did not give much emphasis to this particular group and provide any detailed discussion about it. Instead, the only way to handling this type of data in their studies was to further divide this additional category into several groups on the basis of their semantic meanings.

Based on the above reasons, the present study also adopted the functional classification scheme developed by Biber, Conrad and Cortes (2004) as the analytical framework to characterize the functional use of MWUs in MAIR register. However, different from previous studies, all the extracted domain-specific MWUs in the present study were assigned to each core functional category classified in the taxonomy, as it is believed that the domain-specific MWUs also perform the functions of reference, stance and discourse organizing in the text. For fully understanding the functions served by these MWUs, each category was further divided into several subcategories drawing on the extracted data. The refined functional classification is presented in Table 3.2 below. Finally, an inter-rater reliability was calculated by Cohen's Kappa for the functional classification. The result (0.83) fell within a



satisfactory level of reliability.

**Table 3.2 Functional categories designed for present study**

Functional Categories	Subcategories	Examples
<b>Stance MWUs</b>	a) epistemic stance	<i>is likely to; it is possible</i>
	b) attitudinal/modality stance	<i>crew were unable; were required to</i>
<b>Discourse organizers</b>	a) Topic introduction/focus	<i>noted that; considered that</i>
	b) Topic elaboration/clarification	<i>result in; associated with</i>
<b>Referential MWUs</b>	<b>Identification/focus</b>	
	a) notions	<i>fishing vessels safety; bridge team management</i>
	b) activities	<i>proceeded to; alter course</i>
	c) vessels	<i>a vessel of; national lifeboat</i>
	d) agents	<i>crew members; skipper of</i>
	e) equipments	<i>the starboard engine; the general alarm</i>
	f) regulations	<i>SOLAS chapter; IMO resolution</i>
	<b>Specification of attributes</b>	
	g) Tangible framing	<i>angle of; strength of the</i>
	h) Intangible framing	<i>state of; monitoring of</i>
	<b>Time/place reference</b>	
	i) Time reference	<i>when the vessel was; before the collision</i>
	j) Place reference	<i>the deck of; the fish hold</i>

### 3.2. Research questions

As discussed above, the primary purpose of the present study is to explore the use of domain-specific MWUs in the genre of MAIR. To achieve this research goal, the present study was conducted under the analytical frameworks described in the previous subsection and the following three research questions were addressed:

1. What are the frequency distributions of the domain-specific MWUs in MAIR genre?
2. What are the syntactic features of the domain-specific MWUs in MAIR genre?
3. What functions are the domain-specific MWUs performing in MAIR genre?
4. What are the distinct meanings the domain-specific MWUs possess in MAIR



genre? In what way they deviate from the use in general English register?

As for these four key issues, the first two questions are addressed in Chapter 5, while Chapter 6 provides in-depth discussion of the last two issues.

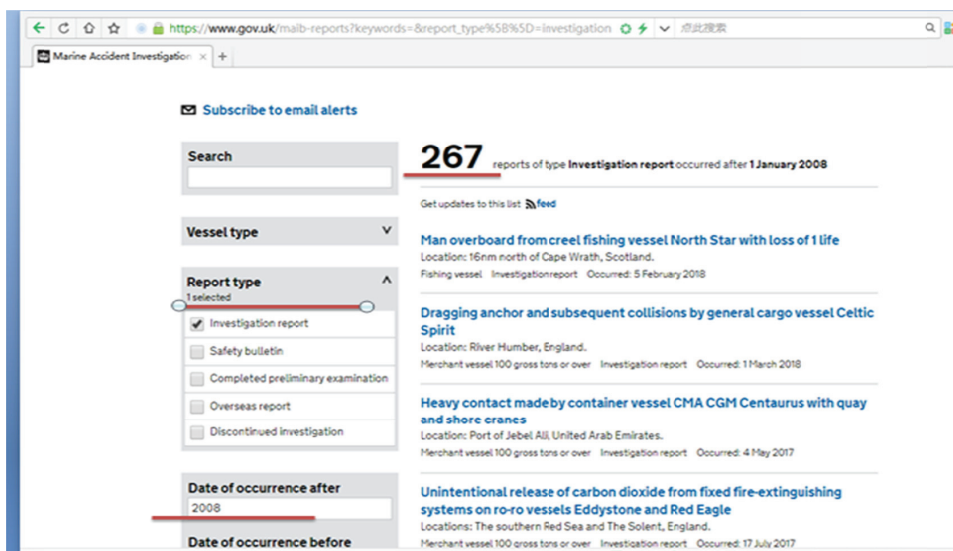
### **3.3. Corpora used in this study**

This section is dedicated to introducing the corpora involved in the present study. First, the compilation of the COMAIR is described. Then it presents a reference corpus used for comparison—British National Corpus Baby (hereafter BNC baby).

#### **3.3.1. Corpus of Marine Accident Investigation Reports (COMAIR)**

The COMAIR is a self-built, domain-specific corpus established for the purpose of the present study. It consists of British marine accident investigation reports ranging from 2009 to 2018, all of which can be freely accessed from the official website of the U.K. Government (<https://www.gov.uk/maib-reports>). Choosing British reports rather than reports of other countries as the target data of the COMAIR is not just because the availability of the data makes it possible to carry out research in this domain. More importantly, the British version of marine accident reports has been commonly recognized as the standardized format in this field. We therefore ensure the representativeness in the selection of data.

The figure below demonstrates how the data was downloaded from the website (<https://www.gov.uk/maib-reports>):



**Figure 3.1 Snapshot of the website where the data was downloaded**

As can be seen from the left side of the website, there is an option of 'report type', which includes five categories: investigation report, safety bulletin, completed preliminary examination, overseas report and discontinued investigation. Among them, the 'investigation report' type was selected as the target data to serve the purpose of the present study. And for the option of 'date of occurrence', also shown on the left side of the website, the time was confined to the period of 'after 2008'. This equals to the investigation reports being published in 2009.

Here, it is important to point out the reason why the study corpus constrain the time period from 2009 to 2018. First, when Bowker and Pearson (2002: 454) provides guidance on how to design and build a Language for specific purposes (LSP corpus), they point out that "you can get more useful information from a corpus that is small but is well designed than from one that is large but is not customized to meet your needs." Their suggestion clearly indicates that exhaustive data is not necessary for building LSP corpus. Rather, the well-structural design of the corpus takes a more important role in it. Moreover, since the present study is to investigate the phraseological patterns of the investigation reports, which is obviously a synchronical investigation. It is thus not necessary to cover all the time periods. By contrast, if this study is a diachronical investigation about the linguistic change of the reports, then it

is better to cover all the time periods. Inspired by the above reasons, the 10 years' investigation reports are believed big enough to represent the characteristics of this genre.

Based on the above criteria, the search results show that there are altogether 267 reports found, as shown in Figure 3.1 above. All these reports were converted into plain text files and cleaned of headings, formatting, diagrams, images and appendices for accurate data processing. Since 12 of the reports failed to do so, hence 255 reports, were finally included in the COMAIR. WordSmith Tools 6.0 software (Scott 2016) was used to profile the COMAIR in terms of word tokens, word types, type/token ratio, word length, etc. All the descriptive data of the COMAIR are outlined in Table 3.3 and Table 3.4 below. As Table 3.4 illustrates, the COMAIR is composed of 1,981,991 word tokens and 23,594 differing types.

**Table 3.3 Descriptive statistics for COMAIR**

Corpus	Covering Years	Number of reports	Total number of words
COMAIR	2009-2018	255	1,981,991

**Table 3.4 Overall statistics of COMAIR**

	Numbers		Numbers
Reports	255	1-letter words	65,336
Ranges of report size	941-43051	2-letter words	312,205
Average report size	7907	3-letter words	440,702
Tokens	1981,991	4-letter words	284,254
Types	23,594	5-letter words	193,310
Type/token ratio	1.19	6-letter words	174,754
Standard type/token	37.93	7-letter words	169,428
Average word length	4.90	8-letter words	136,404
Sentences	82,646	9-letter words	100,612
Sentence length	23.64	10-letter words	65,955
Standard sentence length	18.49	11(+)-letter words	73,228

### 3.3.2. British National Corpus (BNC) Baby

BNC Baby is a four million word sampling of the 100 million word British National Corpus (BNC), a snapshot of British English at the end of 20th century. It was originally developed as a manageable sub-corpus from the BNC for use in the language classroom and was released in October 2004 together with the XML-aware corpus tool Xaira. Similar to BNC, BNC baby comprises samples of spoken and written British English from a wide range of sources but not being restricted to any particular subject field, register or genre, hence has been widely used as a general reference corpus. In BNC Baby, four one-million-word genre-based subsets are included, consisting of academic prose, written fiction, newspaper and conversation (Berglund, Burnard and Wynne 2004).

Considering the size of BNC Baby, which is about twice as large as COMAIR, BNC Baby was chosen as the benchmark in this study to detect the MWUs which are specifically used in the COMAIR. This is because the moderate size of a reference corpus is considered sufficient for KW procedure, according to Scott, who makes a conclusion after investigating various different kinds of reference corpus (Scott 2006, 2009).

Additionally, BNC Baby is designed to be the representative of modern British English, which shares the same language background with the data in COMAIR. By comparisons with BNC Baby, the MWUs representing the common linguistic characteristics of British English in COMAIR will be less likely to stand out. Instead, the MWUs that reflect the specific features of the COMAIR will become particularly noticeable. As claimed by Scott (2009), features which are similar in the reference corpus and the node corpus will not surface in the comparison. Only features where there is a significant departure from the reference corpus norm will become prominent for inspection (Scott 2009: 140). Therefore, choosing BNC Baby as the reference corpus in the present study ensures the validity and representativeness of the comparison results to the most extent. The basic information about BNC Baby is presented in Table 3.5.

**Table 3.5 Descriptive data of BNC Baby**

BNC Baby	Registers	Numbers of texts	Numbers of running words	Percentages
<b>spoken</b>	conversations	30	696,258	17.41%
<b>written</b>	academic texts	30	1,300,467	32.51%
	Fiction	25	1,001,454	25.04%
	Newspapers	97	1,001,821	25.04%
	<b>Subtotal</b>	152	3,303,742	82.59%
<b>Total</b>		182	4,000,000	100%

### 3.4. Tools and procedures for data analysis

The present study combines both quantitative and qualitative analysis. The quantitative analysis aims to detect the amount of MWUs of different structures and functions. By exploring what kinds of structures and functions that the MWUs mostly rely on, the phraseological features of MAIR genre will stand out. While the primary purpose of the qualitative analysis is to understand the use of MWUs from three dimensions corresponding to the analytical framework proposed above. This section is devoted to introducing the detailed procedures of the investigation, in which the tools used to process the data will be described first.

#### 3.4.1. Tools for data processing

##### 3.4.1.1. Wordsmith software

In the present study, both the cluster setting function and keywords function of Wordsmith 6.0 software (Scott 2016) were applied during the process of MWU extraction. The cluster function was adopted to retrieve the n-grams (n=2-5) from the COMAIR while the keywords function was used to generate the keyword list of the COMAIR by comparison with BNC Baby. The way how both functions were applied and assisted in MWU identification will be explicated in detail in Chapter 4. As is known that the cluster function of Wordsmith tools relies on physically adjacent

occurrences of word forms to extract word sequences, but not considering their internal associations. This results in a number of n-grams with no grammatical or semantic status, which doesn't meet the criteria for defining MWUs in our case. Therefore, the outputs of such function can only be treated as the basis for further refinement.

#### **3.4.1.2. R language program**

To further purify the extracted n-grams, a program in R language was developed and ran after n-gram extraction. It was used to calculate the internal association value for each n-gram. The algorithms used in R program are described in Chapter 4. This section only provides brief introduction about the R language program.

R is a programming language and free software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. A wide variety of statistical and graphical techniques are implemented for developing statistical software and data analysis, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering and others. R has stronger object-oriented programming facilities than most statistical computing languages. Firstly, many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made. Another strong point of R is that it is highly extensible through the use of user-submitted packages for specific functions or specific areas of study.

In general, with the tools introduced in this section, the targeted MWUs were extracted from the COMAIR to form the basis of subsequent analysis.

#### **3.4.2. Procedures for data analysis**

The present analysis started with domain-specific MWU identification, an important step determining the validity and reliability of the results. The detailed procedures are introduced in Chapter 4. Afterwards, the analysis turned to assign principal functions and structures to each of the finally identified MWUs. This step

was carried out based on the taxonomy proposed within analytical framework in Section 3.1. Then, the overall frequency was quantified, including the frequency distributions of MWUs across different lengths (e.g., 3-word MWUs, 4-word MWUs, 5-word MWUs), various structures (e.g., NP, VP and PP), and discourse functions (i.e., referential, stance and organizational). The last step of the analysis focused on the in-depth description of the most prominent patterns and semantic meanings of the target MWUs in the COMAIR. Attention was also paid to the variations of these MWUs from their use in general English.

Through this systematic analysis, it is believed to reflect the distinguished phraseological features of the domain-specific MWUs in the COMAIR.

### **3.4.3. Inter-rater reliability**

#### **3.4.3.1. Inter-rater reliability for filtering out process**

The identification of domain-specific MWUs inevitably involves the filtering process. It is operated not only for determining whether the word sequences are qualified as MWU candidates based on the operational definition in the study, but also for screening out the domain-specific sequences from the full list of MWUs. Undoubtedly, this process relies much on researcher's personal judgment which different views would probably occur with. In order to ensure the reliability and validity of the results, one specialist from maritime domain and the author together accomplished these tasks. The Kappa statistic was chosen to measure the agreement between the two researchers, yielding the results of 0.91 and 0.87 respectively. Such high degrees of agreement indicate that the two researchers are highly consistent with reserving or removing certain MWUs. In cases of disagreement, researchers negotiated each case until they reached full agreement. As a result, 1,826 MWUs were regarded as the domain-specific MWUs in the COMAIR.

#### **3.4.3.2. Inter-rater reliability for qualitative analysis**

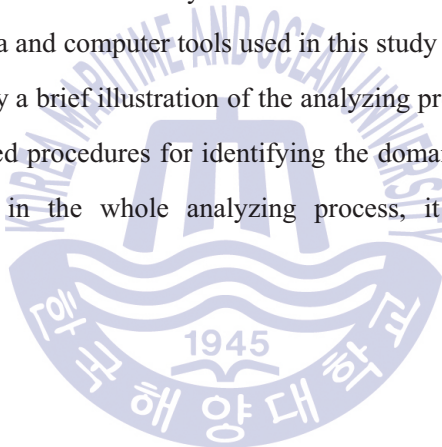
Once the target MWU list was finalized, the last step was the qualitative



investigation of these expressions, including both structural and functional analysis. Again, the structural and functional types of MWUs were manually classified by the two researchers. The ratings of all classifications were aggregated and subjected to statistical analyses in order to assess the inter-rater reliability. The Kappa values in both situations are  $> 0.75$  (0.89 for structural classification and 0.83 for functional classification), which fall within a satisfactory level of reliability. Similarly, researchers discussed each case of disagreement to reach full agreement.

### 3.5. Summary

This chapter has laid a methodological foundation for the current research. It has first proposed a three-dimensional analytical framework and four research questions. Meanwhile, the corpora and computer tools used in this study were introduced as well. It was then followed by a brief illustration of the analyzing process. Next chapter will demonstrate the detailed procedures for identifying the domain-specific MWUs. Due to the crucial status in the whole analyzing process, it thus deserves further elaboration.



## Chapter 4

### Identification of domain-specific MWUs in COMAIR

MWU identification, as the first procedure in any empirical research on phraseology, is not only important for providing a full description of phraseology in a corpus, but also crucial for the validity and reliability of the results. As Stubbs (2005) puts forward, a comprehensive description of phraseology requires a systematic method of extracting the most frequent recurrent strings from the corpus, which can provide evidence for the underlying phrasal units of meanings.

This chapter provides an in-depth discussion of the steps taken to extract, refine and generate the list of target MWUs in the COMAIR.

#### 4.1. Current approaches to MWU extraction

In phraseological studies, a number of statistics-based methods have been developed so far to address the issue of MWU extraction, which mainly falls into two categories: frequency-based measure and measure of collocational association. Despite their applications in various studies, these two approaches have come under criticism for not achieving high precision and efficiency in performance. For example, the frequency-based method, which extracts MWUs by setting the minimum frequency of its occurrences, has been criticized for not always ensure the semantic or functional coherence of the lexical sequences, hence results in a considerable number of 'phraseologically uninteresting' sequences retrieved (Altenberg 199: 133). Beyond that, Simpson-Vlach and Ellis (2010) also points out that the method as such tends to favor lexical sequences which occur often because of their highly frequent individual components, such as function words. Undoubtedly, these two inherent weaknesses lower the precision of the extraction.

The collocational association approach, on the other hand, considers whether the co-occurring words in sequences are meaningfully associated or not, therefore, was

regarded as a breakthrough from former frequency-based algorithm (Church and Hanks 1990; Manning and Schütze 1999; Oakes 1998). It was first introduced by Church and Hanks (1990), who adopted MI statistics to gauge the collocational strength of word pairs in the study of collocations. Later, other measures within this approach, such as Log-likelihood, Z-score,  $X^2$  test, to name a few, have been adopted in various phraseological studies (Devore 2000). However, since this approach is typically applied to measure the salience of the association between two words but not the longer sequences, the currently existing computer software only adopts it to facilitate the identification of two-word sequences. As for sequences longer than two words, frequency-based method is usually applied with a higher priority than the association measure. Furthermore, according to some researchers, this approach does not take into account the order of the words in the sequence, which may be problematic to extract some formulaic sequences, whose formulaicity is partly determined by their fixed word order (Biber 2009).

It thus seems clear from the above that whatever approach is employed, it is not sufficient to produce satisfactory results. To provide a more comprehensive description of the MWUs and their meanings in a corpus, a new approach specifically designed for the present study was proposed.

#### **4.2. My proposed approach to domain-specific MWU extraction**

In order to identify domain-specific MWUs from the COMAIR, the present study proposed a new approach combining statistical-based measure with the the notion of 'keyword'. Reasons for such combination can be illustrated from two perspectives.

First, the use of statistical-based measure instead of frequency-based method is mainly because the domain-specific MWUs tend to occur with relatively low frequency in the specialized corpus. Thus, setting a frequency threshold for identifying domain-specific MWUs inevitably excludes a number of potential target MWUs, which lowers the precision of extraction. By contrast, the statistical-based method puts emphasis on measuring the internal association of the word sequences

and help determine whether the co-occurring words are meaningfully associated or not. Based on this reason, it was applied with higher priority than the frequency-based method for identifying domain-specific MWUs from the COMAIR.

Furthermore, it has been claimed that statistically extracted word sequences are not necessarily MWUs, some of which are even difficult to make sense of. Therefore, the incorporation of linguistic information is needed to improve the extraction performance. A review of relevant literature shows that many phraseological studies incorporate linguistic information for MWU extraction (Daille 1995; Enguehard 1993; Heid 1999; Juesteson 1993). Linguistic knowledge used in these studies covers different levels ranging from semantic field information to syntactic rules and many others, all of which improve the extraction performance to some extent. However, few researchers have approached the MWU extraction from 'meaning' perspective. In fact, it is of significance to do so, especially when attempting to extract the MWUs from a specialized corpus, since phraseology is central and pivotal in meaning creation and language description. It is the MWUs but not the individual words that constitute the basic unit of meaning in the text (Altenberg 1998; Hunston 2002; Pawley and Syder 1983; Sinclair 1991; Teubert 2005; Tognini-Bonelli 2001; Wray 2002). Therefore, we suppose that the 'meaning of text' can be used as an efficient linguistic filter which allows us to extract the MWUs that can best reflect the characteristic meaning of the corpus. Based on this assumption, we proposed the incorporation of 'meaning' into the statistics-based approach for domain-specific MWU extraction. Specifically, we used keywords as the detectors, since it is believed that the domain-specific meanings are conveyed through the MWUs formed around keywords. The choice of keywords as the node words was by no means arbitrary, but principally determined by their significant role in the corpus. As is known, keywords are usually used as an effective and useful method for identifying the discourse topic (aboutness) and stylistic features of texts (Gerbig 2010). This group undoubtedly includes the words which are specifically used within certain domain. And the MWUs around keywords can provide contextual evidence for fully understanding the meaning of these domain-specific keywords. In general, it is supposed that the incorporation of

keywords into statistical method will lead to efficiency improvements in domain-specific MWU extraction.

Table 4.1 below illustrates the procedures of domain-specific MWU identification, in which the proposed approach was applied in steps 2 and 3. The detailed process is demonstrated in the following subsections.

**Table 4.1 Procedures of domain-specific MWU identification**

---

<b>Step 1.</b> N-gram retrieval
<b>Step 2.</b> Keyword-gram extraction
<b>Step 3.</b> Measurement of the association strength of the keyword-grams
<b>Step 4.</b> Filtering out process
<b>Step 5.</b> Domain-specific MWU identification

---

#### **4. 3.The detailed process of domain-specific MWU extraction**

##### **4. 3.1. Step 1: N-gram retrieval**

In light of the operational definition of MWUs in the present study, the initial stage for domain-specific MWU extraction was to retrieve recurrent n-grams ( $n=2-5$ ) from the COMAIR. As mentioned in the previous section, frequency-based measure is not appropriate for achieving the purpose of the present study. Therefore, the n-gram lists were mainly used as the basis for further MWU identification. As for the frequency threshold set for n-gram retrieval, the present study employed 5 as the frequency cutoff point. It is not only because 5 is the default minimum frequency value for extracting n-grams in the WordSmith Tools. More importantly, it has been widely used as the lowest frequency level in most of the phraseological studies, which ensure the recurrent status of the expressions. The entire list of n-grams covering two-, three-, four-, and five-word strings in the COMAIR was generated using the Wordlist program of the WordSmith Tools, as tabulated in Table 4.2. To further narrow the list, a set of keywords was applied in the next step.

**Table 4.2 Distribution of the n-grams of varied lengths (step 1)**

n-grams	Types	Tokens
2-grams	65658	1152780
3-grams	35517	390746
4-grams	11541	103650
5-grams	3203	29290
<b>Total</b>	<b>115919</b>	<b>1676466</b>

#### 4.3.2. Step 2: Keyword-gram extraction

Once the n-gram list was compiled, the next step was to screen out the n-grams which possess the domain-specific nature. This was achieved by applying ‘keywords’ as the basis of the search. In this step, therefore, the extraction started with generating a keyword list by referencing COMAIR against BNC Baby. During this process, the keywords function of WordSmith Tools (Scott 2008) was employed setting the minimum frequency as 3 and Log-likelihood estimate as the statistical measure to decide whether or not an observed frequency difference was significant at the level of  $p < 0.0000001$  (cf. Oakes 1998; Rayson and Garside 2000). Following most keyword analysis, only the keywords with positive keyness value, indicating they are more frequent than expected, were used for the search of domain-specific MWUs. This led to 5204 keywords included in the final list, as shown in Table 4.3.

**Table 4.3 A list of keywords obtained by referencing COMAIR against BNC Baby (step 2)**

Number	Keywords	Keyness value	Frequency
1	vessel	19555.51	12495
2	vessels	8702.21	5646
3	master	8657.14	6161
4	crew	8639.30	5915
5	safety	7657.50	5911
6	port	6938.23	4826

7	skipper	5470.99	3925
8	fishing	5061.20	3455
9	ship	5034.47	3833
10	board	4917.97	4277
.....			
5195	slamming	6.80	16
5196	tack	6.80	16
5197	launched	6.80	104
5198	bold	6.73	52
5199	relieved	6.73	52
5200	slippery	6.67	29
5201	regained	6.67	24
5202	steerage	6.67	24
5203	tiles	6.67	24
5204	underestimated	6.67	24

As can be seen from the table, the majority of the keywords belong to maritime domain conveying the meaning of vessels, people, equipment, etc. This provides evidence to show that keywords can serve as an efficient indicator of the domain-specific n-grams. Once the keyword set was obtained, what followed was using the list as the filter to search the entire list of n-grams. During this process, all the n-grams with which the keywords are embedded were detected and formed a keyword-gram list, prepared for the follow-up refinement. Table 4.4 below illustrates the distribution of the keyword-grams of varied lengths in the list.

**Table 4.4 Distribution of the keyword-grams of varied lengths (step 3)**

Keyword-grams	Types	Tokens
2-word keyword-grams	30278	831646
3-word keyword-grams	26251	432136
4-word keyword-grams	307	26716
5-word keyword-grams	97	10005
<b>Total</b>	<b>56933</b>	<b>1300503</b>

It should be noted here that the way to identify domain-specific n-grams in the



present study is different from previous research, in which the n-gram of this type has been named key cluster or key phrase (Bondi 2010: 3) and has been investigated in many different discourse contexts (Baker 2006; Jhang and Lee 2013a, 2013b; Mahlberg 2007). For example, Jhang and Lee (2013b) conducted systematical investigations within ESP domain to explore the patterns and use of the key clusters in Biomed Corpus and maritime English corpus respectively (Jhang and Lee 2013b).

By looking at the extraction method adopted in these studies, it is not difficult to find that the procedure is the same with the way of keyword detection, which is based on simple verbatim repetition. That means, the status of key clusters is determined by statistical prominence (keyness), calculated through comparing the frequencies of each n-gram in study corpus with its occurrences in reference corpus (Lindquist 2009: 67; Scott 2004; Warren and Greaves 2007). Clearly, this method treats the pattern of co-selection as an indivisible unit instead of placing an equal emphasis on each of the co-occurring words. As a result, there are some cases, where the clusters comprised of important constituents, especially domain-specific keywords, are still overlooked just because they are not statistically outstanding on their own. In these cases, however, it is necessary to treat these clusters as key clusters since they also convey the domain-specific meaning, even though they are not key in a statistical sense.

Therefore, to ensure this class of clusters being extracted, the present study applied keywords to form the basis of key cluster search instead. By shifting the focus away from the entire pattern to the centered keywords that the n-grams consist of, this approach is assumed to extract the domain-specific n-grams to the largest extent.

#### **4.3.3. Step 3: Measuring the association strength of keyword-grams**

##### **The algorithm**

The collocation-based statistical algorithm in the proposed approach involves the concepts of n-gram's dispersion points, pseudo-bigram transformation and fair dispersion point normalization defined by Silva and Lopes (1999).

## N-gram's dispersion points

According to Silva and Lopes (1999), a dispersion point is the space “locate” between the positions of the constituent words of an n-gram. After that point and before it, several other words may appear, showing a kind of “dispersion tendency” at the point. This concept can be illustrated with a recurrent bigram, *chief officer* in COMAIR, which occurs 2,060 times in total and MI score reaches 32.27. The high MI value of this bigram suggests that there is a strong association between the two words. That is to say, when the word *chief* appears in a text, the word *officer* is likely to follow it; and the probability of the word *chief* appearing in the position immediately prior to *officer* is high too. Yet this is not always the case. Because in the COMAIR, it is also possible to find bigrams where the words *chief* and *officer* do not appear together, such as *chief engineer* (occurring 753 times), *chief inspector* (33 times) *second officer* (643 times), *duty officer* (45 times), etc. These instances imply that the bigram *chief officer* has one “dispersion” point, located between the words *chief* and *officer*.

Based on these bigram examples, it can be inferred that every 3-gram ( $w_1, w_2, w_3$ ) has two dispersion points, located between  $w_1$  and  $w_2$ ,  $w_2$  and  $w_3$ , or between  $w_1$ ,  $w_2$  and  $w_3$ . Every 4-gram ( $w_1, w_2, w_3, w_4$ ) has three dispersion points. Every 5-gram ( $w_1, w_2, w_3, w_4, w_5$ ) has four dispersion points. For any n-gram ( $w_1, w_2, w_3, \dots, w_n$ ), there are n-1 dispersion points, with the first dispersion point located after  $w_1$ , the second dispersion point after  $w_2$ , ... the  $(n-1)^{th}$  dispersion point after the word  $w_{(n-1)}$ , as can be illustrated in Figure 4.2 below.

In brief, every n-gram has n-1 dispersion points, no matter what the size an n-gram has.

## Pseudo-bigram transformation

Pseudo-bigram transformation is another concept proposed by Silva and Lopes (1999) to transform every n-gram of size greater than 2 into a pseudo-bigram. That is to say, every n-gram may be seen to have just one dispersion point located between a

left and a right part of the n-gram:  $w_1 \dots w_i$  and  $w_{i+1} \dots w_n$ , where  $i$  can be any value between 1 and  $n-1$  (i.e.,  $1 \leq i \leq n-1$ ). By doing this, it enables us to compare the association values assigned to different size n-grams, and thus study the evolution of the n-gram's association strength when the n-gram's size changes. As Silva and Lopes argues, the information obtained from this evolution is very important for the selection of an n-gram as an MWU.

### **Fair dispersion point normalization**

After transforming every n-gram of size greater than 2 into a pseudo-bigram, the association strength of each n-gram can be calculated. Since there are  $n-1$  dispersion points for each n-gram, then there will be  $n-1$  ways to transform an n-gram into a pseudo-bigram, which produce  $n-1$  different association values for the same n-gram. Thus comes another question: which value can best reflect the whole n-gram's association? Suggested by Silva and Lopes (1999), this problem can be solved by calculating the arithmetic average of the values determined by each dispersion point along the n-gram. It is as if there is a virtual fair dispersion point within the n-gram. In this way, a fair measure of the whole n-gram's association will be obtained.

In combination with the aforementioned three ideas, the collocation-based statistical approach realizes the measurement of the internal association for n-grams longer than two words. With the aid of R language program, this enhanced approach was applied in the present study to extract the MWUs from the COMAIR.

### **Statistical measures**

A review of the relevant literature shows that the statistics-based approach covers a range of statistical methods including Pointwise Mutual Information (hereinafter MI), Log-likelihood ratio (hereinafter LL), Person's Chi-square test (hereinafter  $X^2$ ),  $t$ -test,  $z$ -score test and many others. Clearly, there is no 'best' way of working out association strength for n-grams, as each measure has its own formula to calculate the collocational strength and tends to identify different types of MWUs. As Lindquist

(2009) suggested, the interpretation of collocation data has to take into account which statistical measure has been used. Thus it is necessary to conduct a comparison among these measures and determine which one can best serve the purpose of this study before any further investigations into MWUs in the COMAIR.

## MI

As an information-theoretically motivated metric, MI is probably the most well-known association measure used in corpus-based collocation studies. (Church and Hanks 1989; Church *et al.* 1991) It measures the strength of association between words by calculating the likelihood of two words appearing together within a particular span of words (Biber, Conrad and Reppen, 1998; Church and Hanks 1990: 23). To be specific, it “compares the probability of observing  $x$  and  $y$  together (the joint probability) with the probabilities of observing  $x$  and  $y$  independently (chance). If there is a genuine association between  $x$  and  $y$ , the joint probability [...] will be much larger than chance” (Church and Hanks 1990: 23).

The formula for calculating the MI score is presented below:

$$MI_{(x;y)} = \log_2 \frac{P_{(x;y)}}{P_{(x)} \times P_{(y)}} \quad (4.1)$$

Despite its wide application, MI has been criticized for having a drawback of giving too much prominence to very low-frequency, high contingency combinations, such as the bigrams in which both component words are hapaxes (Biber 2009; Daille 1995; Daudaravičius and Murcinkevičienė 2004: 325-326; Dunning 1993). For the present purpose, which is to explore the most salient phraseological features of COMAIR, these infrequent WMUs extracted by MI are of secondary importance compared to more basic MWUs. Hence, the MI measure was not chosen for MWU discovery in this study.

## **T-test and z-score test**

The calculations of both t-test and z-score test are based on the assumption of the normal distribution of the dataset. As some researchers observe, this assumption is rarely guaranteed in language use unless either enormous corpora are used, or the investigation is restricted to only very common occurrences, such as function words (Church and Mercer 1993; Dunning 1993). As a consequence, it is thought to be problematic to use these two statistical measures if the data are known to be skewed (Dunning 1993). Considering the composition of the COMAIR, which are not normally distributed, it was decided not to take these two measures into consideration.

## **X<sup>2</sup> test**

An alternative test for assessing the dependence of two words which does not assume normally distributed probabilities is the X<sup>2</sup> test. Its calculation is often based on a 2-by-2 contingency table, as seen in Table 4.5. The essence of the test is to examine the extent to which the observed frequencies varies from the frequencies that would be expected if the two words are independent with each other. If the difference is large, the null hypothesis of independence can be rejected, which means that two words depend on each other to form a collocation.

**Table 4.5 A 2-by-2 Contingency Table**

	<b>word<sub>2</sub>: present</b>	<b>word<sub>2</sub>: absent</b>	<b>Totals</b>
<b>word<sub>1</sub>: present</b>	<i>a</i>	<i>b</i>	<i>a+b</i>
<b>word<sub>1</sub>: absent</b>	<i>c</i>	<i>d</i>	<i>c+d</i>
<b>Totals</b>	<i>a+c</i>	<i>b+d</i>	<i>a+b+c+d</i>

Note: the letter *a* and *d* represent the actual (or observed) counts of the cases that the two words  $w_1$  and  $w_2$  co-occur and do not co-occur respectively. Letter *b* refers to the amount of the cases that word<sub>1</sub> occurs but word<sub>2</sub> does not while letter *c* stands for the amount of the cases that word<sub>1</sub> does not occur but word<sub>2</sub> does.

Although it has been used to a wider range of problems in collocation discovery than the two tests described above, the application of the X<sup>2</sup> statistic can be inaccurate

in cases where the expected cell values in the 2-by-2 table are small (Read and Cressie 1988). In other words, it is suggested not using  $X^2$  if the total sample size is smaller than 20 or if it is between 20 and 40 and the expected value in any of the cells is 5 or less (Snedecor and Cochran 1989: 127). Due to the fact that the COMAIR is a specialized corpus containing a number of n-grams with low frequencies, the  $X^2$  test was therefore ruled out in this investigation.

### LL test

For finding sparse data in a corpus, the LL test proposed by Dunning has been empirically proved to be more appropriate and lead to more improved statistical results than  $X^2$  test (Daille 1995; Dunning 1993; Manning and Schütze 2000: 172-175). Furthermore, it does not appear oversensitive to very low frequencies, like the MI does in these cases (Dunning 1993), but allows some frequent MWUs to get onto the list. Therefore, the LL measure has been acknowledged to yield quite good results for multiword extraction. Based on these reasons, the LL test was chosen as the statistical filter to gauge the association strength for each n-gram in the present study. The process of calculating LL score is illustrated below.

Similar to  $X^2$  test, the LL score is calculated on the basis of a contingency table. It adds every cell in the table to the logarithm of that cell and applies the same to multiple combinations of table cells, with the final result multiplied by 2. This entire calculation can be expressed mathematically in (4.2):

$$LL = 2 \times (a \times \log(a) + b \times \log(b) + c \times \log(c) + d \times \log(d) - (a + b) \times \log(a + b) - (a + c) \times \log(a + c) - (b + d) \times \log(b + d) - (c + d) \times \log(c + d) + (a + b + c + d) \log(a + b + c + d)) \quad (4.2)$$

By applying the above three concepts to LL measure, the equation (4.2) can be written as follows:

$$LL_{(w_1 \dots w_n)} = \frac{1}{n-1} \times \sum_{i=1}^{i=n-1} LL_{((w_1 \dots w_i), (w_{i+1} \dots w_n))} \quad (4.3)$$

Clearly, a generalization of the LL formula in this way allows us to find the associative strength of MWUs involving more than two words.

Conventionally, an LL score of 3.84 (95% significance for degrees of freedom 1) is used as the critical value to determine that two items are statistically significant collocates. In the current research, we adopted this criterion. Only the n-grams with the above-threshold LL score are retained and treated as the potential MWUs for further refinements. The results are displayed in table 4.6 below.

**Table 4.6 Distribution of the keyword-grams above LL threshold (LL score  $\geq 3.84$ ) (step 4)**

Keyword-grams	Types	Tokens
2-word keyword-grams	22,324	729,099
3-word keyword-grams	23,720	405,912
4-word keyword-grams	274	24,244
5-word keyword-grams	90	8,984
<b>Total</b>	<b>46,408</b>	<b>1,168,239</b>

By comparing the initial list of the keyword-grams (Table 4.4.) with the refined list by the statistical measure of LL ratio (Table 4.6.), it can be found that 10,525 types of keyword-grams (with the total occurrences of 132,264 times) were discarded due to their below-threshold LL scores, which indicate that their association strength are not statistically significant.

Tables 4.7-4.10 display the first ten keyword-grams extracted with R program, arranged in the descending order of LL value. As can be seen from the tables, the program outputs include not only the extracted keyword-grams and their frequencies, but also the strength of association between the co-occurring words, indicated by the statistical measure of LL ratio. All the keyword-grams tabulated in the tables provide



strong evidence that the statistical filter is an essential technical parameter for MWU identification.

**Table 4.7 First 10 keyword 2-grams arranged in the descending order of LL value**

No.	Keyword 2-grams	LL value	Freq.
1	on board	23906.63	4119
2	chief officer	18006.34	2060
3	would have	17151.05	2459
4	the vessel	14077.21	8611
5	carried out	10842.87	1184
6	fishing vessels	10433.49	1418
7	engine room	10345.26	1146
8	the master	7963.27	4563
9	risk assessment	7660.92	840
10	vhf radio	6869.02	610

**Table 4.8 First 10 keyword 3-grams arranged in the descending order of LL value**

No.	Keyword 3-grams	LL value	Freq.
1	the chief officer	8559.85	1602
2	at the time	7800.80	1165
3	would have been	7175.74	1087
4	the engine room	4941.95	860
5	health and safety	4937.39	401
6	in accordance with	4709.43	744
7	of the accident	4212.69	1331
8	maritime and coastguard	3848.97	318
9	as a result	3618.65	483
10	to ensure that	3344.91	716

**Table 4.9 First 10 keyword 4-grams arranged in the descending order of LL value**

No.	Keyword 4-grams	LL value	Freq.
1	at the time of	5336.61	977
2	maritime and coastguard agency	3552.76	270
3	the maritime and coastguard	2338.23	257
4	prevent similar accidents occurring	2278.44	160
5	determine the contributory causes	2184.30	153
6	basis for making recommendations	2175.87	154
7	contributory causes and circumstances	2164.12	153
8	accident as a basis	2153.88	153
9	analysis is to determine	2135.20	156
10	purpose of the analysis	2115.99	156

**Table 4.10 First 10 keyword 5-grams arranged in the descending order of LL value**

No.	Keyword 5-grams	LL value	Freq.
1	at the time of the	3857.69	859
2	making recommendations to prevent similar	2352.97	154
3	the maritime and coastguard agency	2305.58	216
4	prevent similar accidents occurring in	2301.84	160
5	to prevent similar accidents occurring	2282.23	159
6	contributory causes and circumstances of	2238.84	153
7	recommendations to prevent similar accidents	2233.95	155
8	determine the contributory causes and	2216.45	153
9	the contributory causes and circumstances	2209.76	153
10	to determine the contributory causes	2202.38	153

#### 4.3.4. Step 4: Filtering out process

As statistically extracted n-grams are not necessarily MWUs, some of which are

even difficult to make sense of, it is therefore necessary to undertake a filtering process to further refine the candidate MWU list and narrow the set of MWUs to be investigated. To ensure the accuracy of the exclusion, the establishment of the exclusion criteria took three issues into account: 1) the operational definition of MWUs in the present study, 2) the treatment of the overlapping MWUs and 3) the determination of whether or not the MWUs fall into the ‘domain-specific’ category. Under each circumstance, the exclusion decision had to be made based on human judgment as methodological support (Simpson 2004), since it cannot be automatically classified by the computers. Therefore, the whole process was completed with the assistance of manual intervention. That means the phraseological status of each word sequence in question was determined by manually checking their concordance lines.

In fact, intuitive judgment cannot be completely avoided in phraseological studies (see Altenberg and Eeg-Olofsson 1990; Butler 1997; De Cock, Granger, Leech, and McEnery 1998). As O’Keeffe *et al.* (2007: 79) points out, “although corpus analysis has given us the means to overcome the difficulties involved in the retrieval of MWUs, the automatic retrieval of recurrent strings is only the beginning, and a good deal of inferential analysis is still necessary to see meaning in the lists spewed out by the computer.”

#### **4.3.4.1. Exclusion criteria for the status of MWUs**

As discussed above, the operational definition of MWUs in the present study requires the word sequence constitutes a syntactically related and semantically coherent (or meaningfully associated) unit. Guided by this requirement, any word sequences, which are composed of weakly related syntactic units or do not intuitively look, sound and feel like semantically independent expressions, were considered noise and excluded from the set. This can be exemplified by word sequences such as *the accident the* (532), *accident the* (50), *figure however* (45), *for the vessel to* (43), *vessel when he* (20), etc.

#### 4.3.4.2. Exclusion criteria for overlapping MWUs

In the process of exclusion, the word sequences which are merely fragments of longer units need to be dealt with as well, as the presence of these items has been considered to “inflate the results of quantitative analysis” (Chen and Baker 2010: 33). To achieve this, we checked every keyword-gram from length 5 down to 2 to see whether the shorter word sequences are part of longer phrasal construction, since some shorter multi-word sequences are frequent, only because they are part of recurrent longer-grams (Stubbs 2005). This point is best illustrated by the occurrences of a 2-word sequence *yachting association* and an extended 3-word combination *royal yachting association*. In the COMAIR, both sequences have a similar frequency (57 times), indicating that *yachting association* is derived from *royal yachting association*. In this case, shorter sequences were excluded to avoid unnecessary repetition and make the list as brief and concise as possible. The detection of such overlapping sequences was accomplished by checking each entry in the list of MWUs against with the other entries.

Here it is noteworthy that, when the shorter units occur more frequently than the longer ones, it was decided to preserve them on the list. This is because the shorter sequences in these circumstances function as independent units, which provide additional information about phrases while the longer ones do not. Therefore it is better to treat them separately rather than merge them together. For example, as a part of longer sequence *vessel traffic services* (41), the 2-word sequence *vessel traffic* occur 81 times, which are more frequent than the longer one. Moreover, according to its concordance, the 2-word sequence *vessel traffic* tends to occur as an independent unit in the COMAIR. See the following examples:

- *Despite the density of **vessel traffic** in the area and the proximity to Tower Bridge and HMS Belfast, the company's generic passage plans did not contain details of potential holding areas for its vessels in the event that river piers became temporarily unavailable.*

- *VTS is defined as a service implemented by a Competent Authority, designed to improve the safety and efficiency of **vessel traffic** and to protect the environment.*

#### 4.3.4.3. Exclusion criteria for domain-specific MWUs

As the present study targets at the domain-specific MWUs in MAIR genre, it is therefore of significance to clarify the meaning of this term. In general, the domain-specific category involves two types of MWUs. One is the MWUs containing domain-specific words (e.g., *vessel*, *master*, *anchor*, *starboard*, etc.) The MWUs of this type are usually technical terms and expressions which convey specialised meaning in maritime domain. Thus they were undoubtedly treated as the target of the present study. Another type consists of MWUs, which possess general nature but perform pragmatic function in MAIR genre. The inclusion of such MWUs in domain-specific group is mainly because they are highly conventionalized in certain text. To be specific, their meaning is shaped by the interplay of linguistic and extra-linguistic factors and their use is tied to standardized communicative situations (Coulmas 1981; Erman and Warren 2000; Eskes 1997, 1999, 2003). Clearly, of all the MWUs with various functions, both stance and discourse organizing MWUs tend to be pragmatically-loaded, since their use is restricted by the situation in which they occur. However, it is not the case for referential MWUs.

For the above reasons, in order to be qualified as a domain-specific MWU, the candidate word sequence should be either composed of maritime-specific words or general by its nature but perform pragmatic function (i.e., stance and discourse organizing functions) in the MAIR genre.

It is worth noting that, this process relies much on researcher's personal judgment. That is, although most domain-specific MWUs can be easily detected, there are still some keyword-grams, which the literal meanings provide an indication of general nature, but they are indeed used in their terminological sense in the COMAIR. This can be exemplified by the keyword-gram *a passage* (55). Closer inspection of the

concordances in the COMAIR showed that this expression specifically serves to convey the meaning of the journey or movement of the vessel, as shown in the following sentences. It was thus decided to be maintained in the final list.

Accordingly, a check of the concordance lines is needed in any case of keyword-gram in question.

N	Concordance
1	tidal stream data in the area increases the difficulty in planning and executing a passage into and from Umm Al Qaywayn. With an axis of 167°/347° between the lateral
2	described in MGN 538, it would have been prudent and best practice to have developed one. A passage plan not only sets out the intended route, but it also identifies navigational
3	Code also required all commercial vessels to conduct operations in accordance with a passage plan that should take into account all obstructions on the route. The passage
4	all relevant information required for the intended passagePlanning' Developing and approving a passage plan based on the outcome of the appraisal of all relevant informationExecution'
5	planThe ICS guide talked about planning for the ocean, coastal and pilotage phases of a passage. It acknowledged that it might be impractical to include all details in the passage
6	12nm off in order to avoid any recreational or coastal fishing vessels. However, planning a passage that at times passed close to the coastline left limited sea room for the tug and
7	ALP Forward's master mitigate the effects the heavy weather had on the tow line. Planning a passage so close to the coastline left limited sea room for the tug and tow to drift in the
8	identified in the report as mitigating this risk. 1.9.3 Sailing directions For vessels planning a passage through the Pentland Firth, Admiralty Sailing Directions (North Coast of Scotland
9	8 hours berthed in Rordal for each loading operation and the charter agreement required a passage speed of about 9 knots's. On 30 December 2014, Aalborg Portland's cement
10	Rordal, Denmark to Rancorn, UK, the most direct route is 981nm via the Pentland Firth. A passage via the English Channel, avoiding the Pentland Firth, is 1,187nm (Figure 36). On
11	unlikely that the master would have considered this option. Cemford's charter required a passage speed of about 9 knots's, in good weather, which was only slightly less than the
12	stream is opposed by strong W or NW winds.'s Identifying the pre-conditions for aborting a passage or taking an alternative route should form part of the navigational plan. A decision
13	and a deckhand. At 0100 (29 August 2015), with the vessel proceeding on autopilot at a passage speed of 9.5kt, the skipper handed the bridge watch over to the deckhand and
14	on the requirement to make an appraisal of potential navigational hazards when planning a passage. A high concentration of fishing vessels in coastal waters presents a very
15	skipper. His ocean experience included his Yachtmaster Ocean qualifying trip as mate on a passage from the Azores to Cork, and as skipper of Cheeki Rafiki on the ARC 2013
16	a chart or uploading an electronic file onto an electronic chart system. The safe execution of a passage relies on the robust appraisal of all information relevant to the proposed voyage,
17	safety corridor. Particularly in pilotage waters, the XTD should be calculated for each leg of a passage and take into account the expected width of safe water available. For the leg of the
18	appropriate, updated charts been available on board, Isamar's master could have prepared a passage plan which would have enabled him to ensure that the intended route was suitable
19	safe option available on board for passage planning and monitoring. The master's approval of a passage plan based on the ECS and without reference to the appropriate paper charts
20	still being developed. As a result, paper charts were the primary means of navigation. Once a passage plan had been approved by the master, the intended track (including a 5c safety

**Figure 4.1 First 20 concordance lines of a passage in COMAIR**

In order to ensure the validity of the final list, the task was completed by consulting a specialist from maritime domain, since understanding domain-specific vocabulary requires a certain degree of scientific knowledge, and teaching them is usually the role of specialists in the field, not of language teachers (Nation 2001). The inter rater reliability was tested via kappa statistic to measure the agreement between the two researchers and the results were presented in Chapter 3 above.

Finally, word sequences composed of personal and place names were excluded from the count as they do not hold interesting meanings or functions, such as *city of Rotterdam* (80), *the united kingdom* (68), etc.

#### 4.3.5. Step 5: Domain-specific MWU identification

With the application of the above exclusion criteria, word sequences were finally



selected to represent the domain-specific MWUs in the COMAIR (Table 4.11.).

**Table 4.11 Distribution of the domain-specific MWUs of differing lengths (step 5)**

Domain-specific MWUs	Types	Tokens
2-word MWUs	9238	374586
3-word MWUs	6008	139038
4-word MWUs	971	17932
5-word MWUs	253	4099
<b>Total</b>	<b>16470</b>	<b>535655</b>

For the qualitative analysis of these MWUs, the list has to be narrowed down to a more manageable size. Therefore, only the ones occurring more than 40 times were targeted, since such threshold level has been proved to provide sufficient number of recurring expressions, thus have the most potential to yield interesting results. (Cortes 2002; Simpson 2004) Altogether, there are 1,826 MWUs in the list for subsequent qualitative analysis (Table 4.12.). Appendix 1 lists all the 1,826 MWUs together with their structural and functional types.

**Table 4.12 A list of domain-specific MWUs (frequency $\geq$ 40)**

Domain-specific MWUs	Types	Tokens
2-word MWUs	1111	160217
3-word MWUs	594	59469
4-word MWUs	96	7670
5-word MWUs	25	1753
<b>Total</b>	<b>1826</b>	<b>229109</b>



## Chapter 5

### Frequency Distributions and Syntactic Features of domain-specific MWUs

This chapter is dedicated to addressing the first two research questions of the present study. That is what the frequency distributions and syntactic features of the domain-specific MWUs are in MAIR genre.

Before that, it is important to consider the type-token distinction. Since there are some cases that one category is assigned to a plenty of different MWU types but each occurs with low frequency. Sometimes, the reverse situation can also happen: despite of few types of MWUs included in the category, there is a large number of occurrences for each type. Therefore, when comparing distributions across different categories, frequency counts are provided for both MWU types and tokens.

#### 5.1 Frequency distributions of domain-specific MWUs

To understand the frequency distributions of the target MWUs in the COMAIR, distributions of MWUs in various lengths and across different frequency bands are discussed respectively.

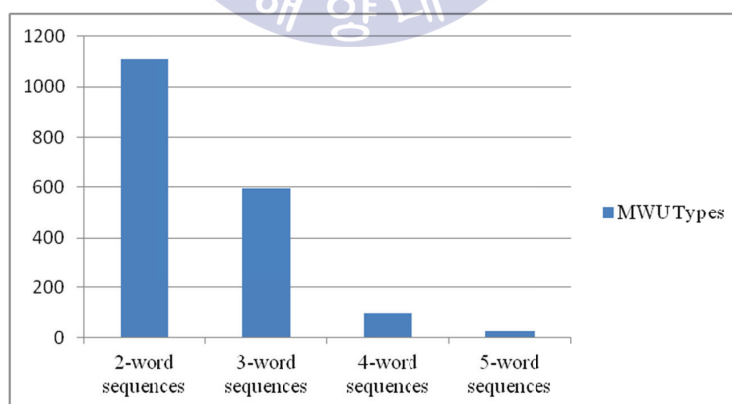
##### 5.1.1. Frequency distributions of the domain-specific MWUs in various lengths

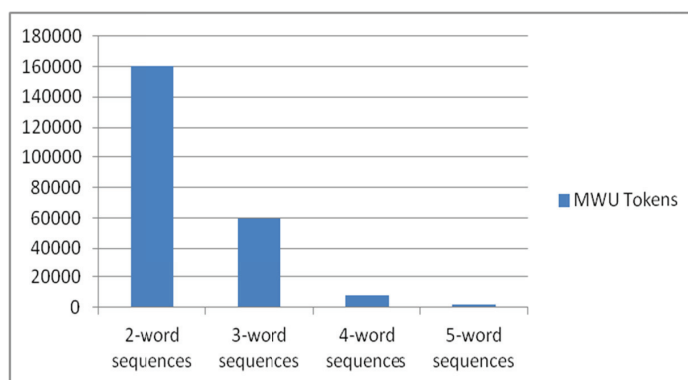
After applying the exclusion criteria, 1,826 word sequences (henceforth types) remained on the final WMU list, with lengths varying from two to five words. These WMUs amounting to a total of 229,109 individual instances (henceforth tokens) account for 11.46% of the roughly two million words in the COMAIR. Table 5.1 presents the frequency distributions of the domain-specific MWUs of different lengths within the list.

**Table 5.1 Distribution of the domain-specific MWUs of differing lengths (frequency $\geq$ 40)**

MWUs	Types	Type percentage (%)	Tokens	Token percentage (%)
2-word sequences	1111	60.84%	160217	69.93%
3-word sequences	594	32.53%	59469	25.95%
4-word sequences	96	5.26%	7670	3.35%
5-word sequences	25	1.37%	1753	0.77%
<b>Total</b>	<b>1826</b>	<b>100%</b>	<b>229109</b>	<b>100%</b>

As can be seen from Table 5.1, the list is largely composed of two-word sequences, whether in terms of types (60.84%) or tokens (69.93%). They are followed by three-word sequences, which make up 32.53% of the MWU types and 25.95% of the total tokens. The occurrences of 4- and 5-word MWUs are relatively rare in the COMAIR, having proportions at 5.26% and 1.37% respectively among all the domain-specific MWU types. Equally, these two groups of MWUs reduce substantially in tokens as well, with just 3.35% and 0.77% in each case. Figures 5.1 and 5.2 below are the graphic representations of the distributions of the domain-specific MWUs in various lengths.

**Figure 5.1 Types of domain-specific WMUs of various lengths in COMAIR**



**Figure 5.2 Tokens of domain-specific MWUs of various lengths in COMAIR**

Both figures display an inverse relation between the length and frequency of MWUs. That means, with the increase of the sequence length, the frequency of MWUs decreases substantially, which finally achieve a diminishing status. This result indicates that the use of domain-specific MWUs also displays unique and creative feature as phraseology does in general English (Coulthard 2004; Sinclair 1987, 2001). As some researchers claim, the prevalence of phraseology in the language does not mean that language use is not unique or creative (Coulthard 2004; Sinclair 1987, 2001). In fact, “even a sequence as short as ten running words has a very high chance of being a unique occurrence” (Coulthard and Johnson 2007:198).

In this study, reason behind such phenomenon can also be explained by the decision of excluding the MWUs, which are syntactically irrelevant and semantically incoherent units. This undoubtedly reduces the number of four- and five- word target MWUs.

### **5.1.2. Overall frequency distribution across different frequency bands**

Apart from the above findings, another observation was made to the overall distribution of the domain-specific MWUs across different frequency bands. Table 5.2 displays the results. As can be seen from the table, not so many domain-specific MWUs are actually common, occurring more than 1,000 times in the COMAIR. In

fact, only 20 MWUs (1.10%) reach this frequency level while 1296 MWUs (70.97%) occur with the frequency less than 100 times. This skewed distribution holds the point that MAIR tends to employ a wide variety of domain-specific MWUs rather than repetition of a small number of common expressions. One possible reason is that the content of each MAIR is of its nature so diverse and extensive that few domain-specific MWUs occur very commonly within the genre. It may also be due to the ESP-based nature of this genre, which brings about many technical terms and idiosyncratic expressions with low frequency.

**Table 5.2 Distribution of the domain-specific MWUs across frequency bands**  
(frequency $\geq$ 40)

Frequency band	Number of MWUs	percentages
$\geq 1000$	20	1.10%
500-999	24	1.31%
100-499	486	26.62%
$\leq 99$	1296	70.97%
<b>Total</b>	<b>1826</b>	<b>100%</b>

Table 5.2 tabulates all the domain-specific MWUs occurring at least 1,000 times in the COMAIR. As table shows, altogether 20 word sequences are included in the list, most of which employ NP-base structure and fall into two main semantic categories: vessel and people. Among them, 10 represent vessels (e.g., *the vessel*, *the ship*, *fishing vessels*, etc.) or parts of the vessels (e.g., *the bridge*, *engine room*, *the deck*, etc.). The rest six MWUs refer to the people working on board the vessel (e.g., *the master*, *the chief officer*, *the pilot*, etc.). Since the genre being investigated is the report of marine accidents, in which both human and vessel play inevitable and essential roles, it is hardly surprising to find the MWUs of these two types top the list. However, of all these expressions, the MWUs *on board* and *would have been* become striking since both of them neither belong to NP-based structural category

nor refer to vessel and people. For *on board*, it has the 3<sup>rd</sup> highest frequency (4,119 times) in the COMAIR and is usually used to specify the activities carried out on the vessel. Therefore, it can be said that specification is one of the distinguished features of MAIR genre. While the MWU *would have been* occurs 1,087 times in total ranking as the 15<sup>th</sup> of the list. As is known, this MWU is commonly used in the subjunctive mood to establish a hypothetical or possible situation. It thus can be inferred that through highly frequent use of *would have been*, the reporters express the wish that the certain acts would be done and accidents would not happen.

**Table 5.3 A list of domain-specific MWUs occurring over 1000 times in COMAIR**

Rank	MWUs	LL value	Frequency
1	the vessel	14077.21	8611
2	the master	7963.27	4563
3	on board	23906.63	4119
4	the skipper	4814.09	2810
5	the crew	2961.44	2625
6	the port	2722.32	2269
7	the ship	3037.11	2147
8	the bridge	3176.47	2112
9	chief officer	18006.34	2060
10	the engine	2491.61	1829
11	the chief officer	8559.85	1602
12	fishing vessels	10433.49	1418
13	of the vessel	1488.0478	1182
14	engine room	10345.2606	1146
15	would have been	7175.7438	1087
16	the pilot	1387.2955	1082
17	the starboard	1328.4991	1079
18	the deck	653.6835	1022
19	the boat	1219.4652	1020
20	the wheelhouse	1932.3584	1015

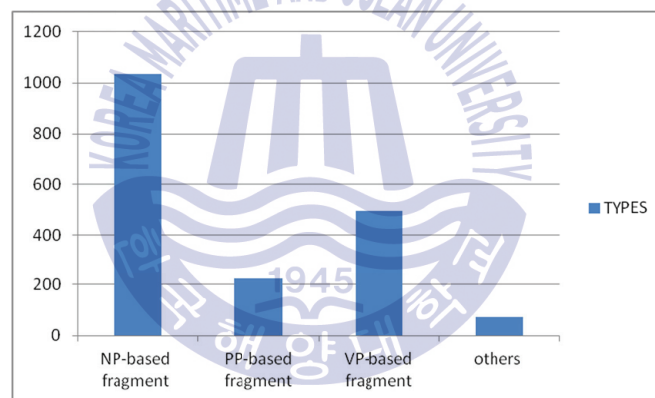
## 5.2. Syntactic features of domain-specific MWUs

To understand the syntactic features of the domain-specific MWUs, all the target MWUs are classified into the structural taxonomy, in which four broad groups are included: VP-based pattern NP-based pattern, PP-based pattern and others. Table 5.4

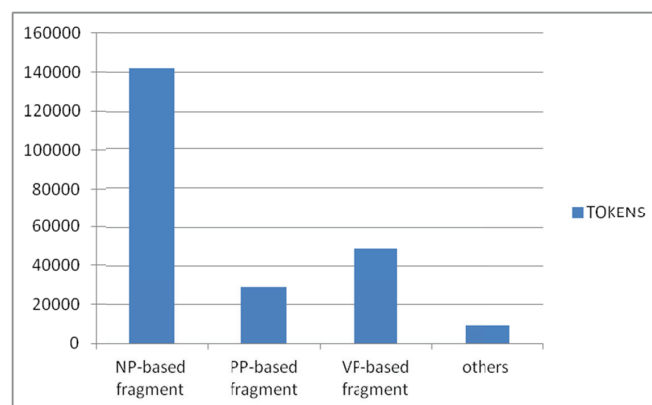
shows the structural classification of the MWUs with the corresponding types, tokens and their percentages. Figures 5.3 and 5.4 display the distributions of the MWUs of different structural types and tokens.

**Table 5.4 Structural classifications of the domain-specific MWUs in COMAIR (frequency $\geq$ 40)**

Structural types	Types	%	Tokens	%
<b>NP-based structure</b>	1035	56.68%	141930	61.95%
<b>PP-based structure</b>	223	12.21%	28972	12.65%
<b>VP-based structure</b>	495	27.11%	49032	21.40%
<b>Others</b>	73	4.00%	9175	4.00%
<b>TOTAL</b>	1826	100%	229109	100%



**Figure 5.3 Distribution of structural types**



**Figure 5.4 Distribution of structural tokens**

The above table and figures demonstrate that NP-based structure is the most common grammatical type among all of the target MWUs, accounting for over 56.68% of MWU types and 61.95% of tokens. The abundant use of NP-based fragment implies that the domain-specific meaning of MAIR genre is largely carried in nominal group. In other words, by making direct reference to maritime-related entities, these nominal MWUs contribute to informational or propositional meanings in the COMAIR. This result seems reasonable in that the highly specialized nature of COMAIR results in a large number of technical terms included. These technical terms carry a broad range of specialized meanings in maritime domain including concepts (e.g., *emergency procedures*, *construction standards*, *intended track*), names of entities (e.g., *radar display*, *container ship*, *engineer deckhand*), regulations or organizations (e.g., *international maritime organization*, *equipment regulations*), and so on. Thus, they provide ample evidence for understanding the meaning construction of the domain-specific MWUs in MAIR genre. Apart from technical terms, the NP-based group also includes a number of MWUs constituting the pattern of ‘NP + of’. This pattern is usually used to single out some particular attributes of an entity (e.g., *an angle of*, *surface of*, etc.); to specify processes or actions (e.g., *discharge of*, *the installation of*, *maintenance of*, etc.); and to make direct reference to agents and locations (*masters of*, *starboard side of*).

As Table 5.4 indicates, there is also a marked prevalence of VP-based structures among the domain-specific MWUs in the COMAIR (27.11% of MWU types and 21.40% of tokens). These MWUs present structural variation, such as copula *be* + PP fragment, *to*-clause, *that*-clause fragment, etc. Of all these subtypes, the ‘verb phrase with active verb’ pattern stands out since it incorporates a large number of action verbs. Close examination of these phrases found that the action verbs are especially used to refer to the actions done by the people. Instances include *left the bridge* (77), *had worked on board* (70), *arrived on the bridge* (66), *wearing a lifejacket* (66), *fell overboard* (53), *informed the master* (43), etc. The wide use of these phrases implies that the MAIR genre tends to highlight the people’s roles during the accidents, with particular attention to the information about what or who caused or performed the



activity. This result appears somewhat surprising since in formal writings such as academic papers or official reports, actions are usually considered more significant than the agents of the actions (*Oxford Dictionaries* 2017). For this reason, this finding reflects the characteristics of the MAIR genre, which is distinguished from other formal writings.

PP structures were also commonly employed by domain-specific MWUs for meaning realization, especially those beginning with *of* (i.e., MWU types and tokens) account for around 40%, while the MWUs starting with other prepositions, such as *from*, *in*, *on* to name but a few, altogether take the percentage of 60%. The frequent and varied use of these patterns is mainly to specify possessions and relationships. It can thus be inferred that the domain-related information that provided in MAIR genre tends to be concrete and specific. Evidence can be seen from the MWUs *of small fishing vessels*, *of deck*, *of the engine room*, *of the bridge*, *of the pilot*, etc.



## Chapter 6

### Functional and Semantic Features of domain-specific MWUs

This chapter concentrates on the meaning analysis of the target MWUs, which include both functional and semantic aspects. The functional interpretation is intended to reveal the specific functions that the domain-specific MWUs perform in the COMAIR and the typical patterns they employ to realize these functions, while the semantic analysis aims to uncover the distinct meanings the domain-specific MWUs possess in the COMAIR, with a particular interest in exploring any variations from the use in general English register. As mentioned in Chapter 3, domain-specific MWUs include a number of expressions, which are of general nature on their surface, but are used largely in their terminological sense in the COMAIR, such as *strength of*, *state of*, etc. Therefore, these expressions lie in the focus of the semantic analysis in the present study. To better understand the semantic meaning of this type of MWUs, a comparison with general English, in this case, the BNC Baby, was undertaken. It helped provide evidence to show whether there exist any semantic differences in the use of these MWUs in each register and what their distinctive semantic features in the COMAIR are. Since the functional and semantic meaning are somewhat inseparable (Lakoff 1987), this chapter will discuss these two aspects at the same time.

#### 6.1. Distributions across primary discourse functions

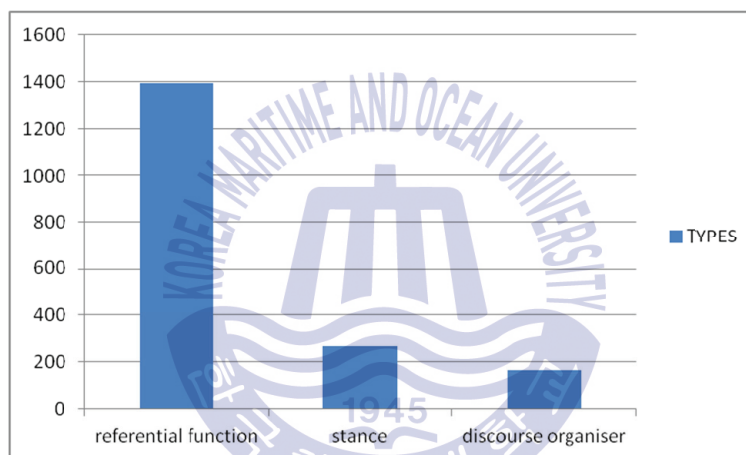
In the functional classification scheme proposed by Biber *et al.* (1999), the domain-specific MWUs were assigned to one of the three functional categories based on their typical meanings and uses. Thereby, the extent to which each functional category is used in the COMAIR can be revealed. It is believed that the functional analysis helps gain a better awareness of the particular concerns of this type of discourse.

Table 6.1. shows the frequency distributions of the MWU types and tokens across

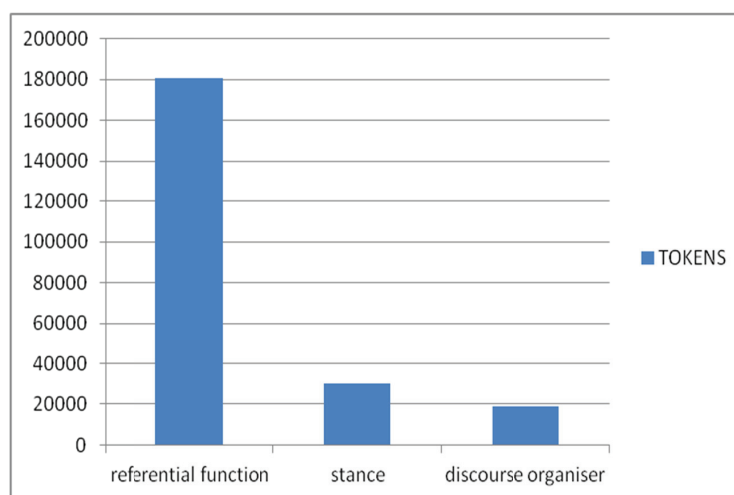
these primary functions in the COMAIR.

**Table 6.1 Frequency distributions across primary discourse functions (frequency $\geq$ 40)**

Function	Type	%	Token	%
Referential MWUs	1396	76.45%	180453	78.76%
Stance MWUs	266	14.57%	29920	13.06%
Discourse organizer	164	8.98%	18736	8.18%
<b>Total</b>	<b>1826</b>	<b>100%</b>	<b>229109</b>	<b>100%</b>



**Figure 6.1 Distribution of functional types**



**Figure 6.2 Distribution of functional tokens**

It is clear from the above table and figures that referential MWUs is among the commonest functional types in the COMAIR, making up the dominant proportion at 76.45% and 78.76% in terms of types and tokens. This result demonstrates that the primary function of domain-specific MWUs is to express referential meanings. Semantically, these referential MWUs can be classified into several types, i.e., identifying activities or actions, specifying entities, notions and attributes, expressing space, etc. Although each subtype has its own forms and features, they all contribute to the thematic development of MAIR genre propositionally.

Apart from referential MWUs, there are a relatively small number of domain-specific MWUs serving the function of stance in the COMAIR, accounting for 14.57% of types and 13.06% of tokens. Within this group, most MWUs express impersonal epistemic stance, reflecting the stance feature of MAIR genre, where objectivity is essentially required. Other word sequences appear to be deontic in nature, the primary function of which is to set out the obligations and issue suggestions for the agents.

By contrast, domain-specific MWUs functioning as discourse organizers have the least number of both types (8.98%) and tokens (8.18%). This group usually adopts typical patterns to serve the discourse function of topic introduction and clarification. For example, the function of topic introduction is mostly achieved by the use of *that* clause patterns, such as *established that*, *there is no evidence to suggest that*, *it was reported that*, etc.

## 6.2. Multiple functioning

Among the target MWUs on the list, a bunch of expressions (nearly 30% of the MWU types) were found to be multifunctional. Their occurrences can be explained by the decision to include only the domain-specific MWUs in the present study. As discussed in Chapter 3 above, the domain-specific MWUs are the word sequences contributing to the meaning of the text. Therefore, these expressions primarily fulfill

referential functions in MAIR genre. However, in some cases, these referential MWUs also serve other functions depending on the different co-texts, which lead to the presence of multi-functionality. For example, some MWUs possess secondary stance functions, typically as modalized epistemic or deontic markers. Evidence can be found from the MWU *the master should*. On one hand, this sequence makes direct reference to *master*; hence presents propositional information about this agent. On the other hand, this MWU clearly expresses the stance, as the modal verb ‘*should*’ in this sequence points out the obligation of the master for performing the act suggested by the ensuing proposition. This can be demonstrated by the following instances.

- Before commencing a tow, *the master should* determine which towing gear is suitable for the operation and instruct the crew accordingly.
- *The master should* exercise prudence and good seamanship having regard to the season of the year.

Apart from the above situation, it is sometimes difficult to distinguish organizational and referential functions. One of the typical examples is the MWU *master decided to* (44). From referential perspective, this expression functions to put emphasis on the master’s performance of ‘decision making’. However, from organizational perspective, the MWU *master decided to* serves as a signal of topic elaboration, with the purpose of providing additional explanation or clarification about the content of the decision, as shown below.

- *The master decided to* remain on the bridge until the tug and tow were once again moving ahead.
- *The master decided to* proceed to Tobermory at the best possible speed so as to be at anchor in the shelter of Tobermory Bay before the weather deteriorated.
- ALP Forward's *master decided to* use this as another opportunity to manoeuvre the tug in an attempt to influence the direction of drift

away from the approaching coast.

For the issue of multi-functionality, a solution is to examine the concordance lines of the potentially multifunctional MWUs and assign one salient function according to their most common use and meaning in different contexts (Biber *et al.* 2004). This approach is feasible in the present study since the previous filtering process narrow down the target MWU list to a manageable size. Hence it was adopted to treat the MWUs which serve varied functions in the COMAIR.

### **6.3. Stance MWUs**

This section focuses on the functions and meanings of stance MWUs in MAIR genre, with an attempt to characterize their most prominent features. Discussion starts with clarifying the notion of stance MWUs in light of Biber *et al.* (1999, 2006). It is then followed by the detailed investigation into this functional category from form and meaning perspectives.

#### **6.3.1. Notion of stance MWUs**

According to Biber *et al.* (1999), stance MWUs provide a typical frame of ‘attributes, value judgments, or assessments for interpreting a propositional content or explicitly addressing readers to draw their attention or influence them’ (Biber *et al.* 1999: 966). That means, stance MWUs are usually used to evaluate or comment on certain performance or behavior and to voice viewpoints, take stance, and offer suggestions to others (Biber, *et al.* 2006).

#### **6.3.2. Stance MWUs in COMAIR**

In MAIR genre, a number of MWUs are used to realize the purpose as such. For example, in the study of lexical bundles in ESP writing by Jhang *et al.* (2018), the authors found that impersonal epistemic bundles, such as *it is likely that*, are preferably used in MAIR genre when native English reporters draw inferences about

the causes of the accidents. These lexical bundles reflect the degree of uncertainty that the reporters hold towards the investigation results. Although Jhang *et al.*'s (2018) study is exploratory in that it discovered the typical patterns of stance bundles in MAIR genre, their findings only focus on the most frequently occurred lexical bundles, which are often the expressions with general nature. The stance MWUs with domain specific nature were not touched upon. In fact, by close look at the domain-specific MWUs in MAIR genre, the present study found that domain-specific MWUs serve a wider range of stance in MAIR genre. For instance, some MWUs tend to convey the meaning of necessity for performing certain acts (e.g., *should be taken*, *was required to*, etc.) and the ability that the agents own or need (e.g., *crew were able*, *crew were unable*, *vessel could*, etc.). Table 6.2 demonstrates the frequency distributions across subcategories within stance functions.

**Table 6.2 Frequency distributions across subcategories of stance function (frequency $\geq$ 40)**

Subtypes	types	%	tokens	%
Epistemic stance	167	63.74%	17406	62.12%
attitudinal/ modality stance	95	36.26%	10616	37.88%
<b>Total</b>	262	100%	28022	100%

As can be seen from Table 6.2, around 60% of the domain-specific MWUs express epistemic stance in MAIR genre while the rest 30% of word sequences serve the function of attitudinal/modality stance. Further scrutiny of the epistemic stance group also allowed us to find that all the epistemic MWUs are impersonal, with the purpose of minimizing the imposition of the reporters' opinions. Moreover, these expressions tend to employ the pattern of 'copula *be* + adj.' to express the degrees of both certainty and uncertainty. Such findings were obtained from the high frequencies of this pattern when realizing epistemic function (Figure 6.3), and the adjectives that occur within the pattern, as shown in the Table 6.3 below.



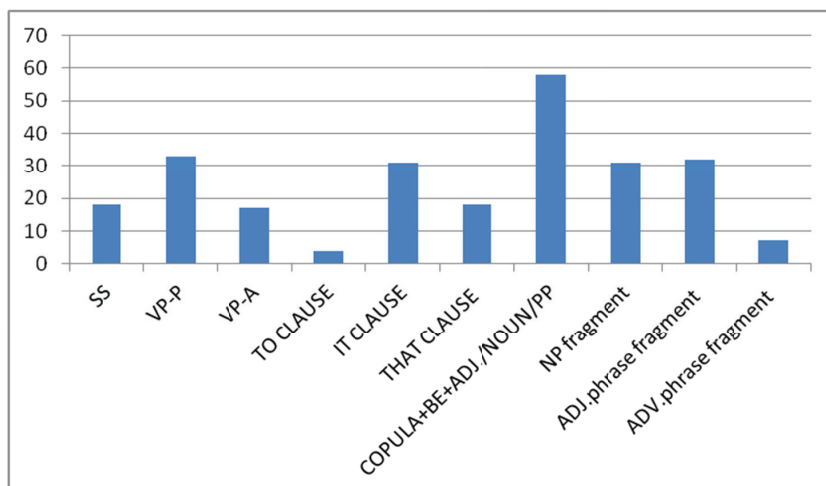


Figure 6.3 Distributions of the syntactic patterns serving the function of epistemic stance

Table 6.3 Adjectives used in the pattern of ‘copula *be* + adj.’

Pattern	Adjectives
‘copula <i>be</i> +adj.’	possible <sub>(1022)</sub> , likely <sub>(733)</sub> , aware <sub>(628)</sub> , unable <sub>(373)</sub> , clear <sub>(261)</sub> , responsible <sub>(256)</sub> , apparent <sub>(136)</sub> , evident <sub>(295)</sub> , unlikely <sub>(151)</sub> , essential <sub>(121)</sub> , safe <sub>(118)</sub> , probable <sub>(70)</sub> , capable <sub>(43)</sub> ,

Unlike epistemic stance, the attitudinal/modality stance is mainly realized by the MWUs incorporating with *require* or modal verbs, as demonstrated in the following examples.

- MSC/Circ.884 also states that the continuous bollard pull of the towing **vessel should** be sufficient to maintain control of the towed vessel in the following environmental conditions.
- Antari’s MSMC was issued by the Antigua and Barbuda administration on 18 January 2008. It specified that a minimum of six **crew were required** for the vessel, consisting: master, chief officer, chief engineer, two ratings forming part of a navigational watch and one deck rating.

As can be seen from the above examples, these MWUs are used to set out the obligations and issue suggestions for the agents. Specifically, they serve to direct the

agents that some actions ought to be carried out according to certain norms and regulations rather than reporters' personal expectations or desire. This can be found from many cases where regulations or rules precede these expressions. Thus, it can be argued that the obligation MWUs often shunt back and forth between real world events and laws and tend to establish a relation between these two. By disclosing the source of norms, it enables the reporters to take an objective stance.

Among all the attitudinal/modality MWUs, the word sequence *would have been* stands out, since it is employed as the most usual one to express stance, with occurrences of 1,087 times in the COMAIR. Table 6.4 tabulates the list of stance MWUs arranged in the descending order.

**Table 6.4 A list of stance MWUs arranged in the descending order (frequency $\geq$ 40)**

Number	Stance MWUs	LL value	Frequency
1	would have been	7175.744	1087
2	would be	2736.991	646
3	unable to	1990.812	481
4	aware of	1624.85	449
5	likely that	2101.761	410
6	required by	1313.694	360
7	intended to	940.0118	344
8	was required	583.3373	333
9	is likely to	2004.173	333
10	possible that	1499.492	331
.....			
260	it was clear	240.4696	40
261	it was safe to	229.9501	40
262	was possibly	106.8618	40
263	was unlikely	96.0259	40
264	been possible to	170.1003	40
265	more likely to	216.5258	40

By further checking the concordance lines of this expression, it is found that *would have been* is commonly used in the subjunctive mood to establish a hypothetical or possible situation, through which reporters express the wish and willingness that the certain acts would be done and accidents would not happen. For example,

- Had they been wearing PFDs and been able to raise an alarm, their chances of survival **would have been** significantly improved.
- To do this, he **would have been** facing aft and therefore unable to see the lights being shone by the occupants of James.
- The level of emissions **would have been** even higher at faster engine speeds, particularly as the engine was 16 years old.

In some occasions, *would have been* is also used to express the certainty or likelihood that something was the case in the past. When serving this function, expressions of this type tend to collocate with general MWUs, which also convey the meaning of certainty or likelihood, such as *it is likely that*, *it is very unlikely that*, etc, as exemplified by the following sentences.

- Had Scott been wearing a lifejacket when he entered the water, it is likely that he **would have been** recovered alive.
- Had he considered the weather off the north-west coast of Scotland in his planning, it is unlikely that he **would have been** concerned.
- Had the wave rider buoy been visible and clear of the rig, it is by no means certain that it **would have been** possible for any vessel to safely recover it in the severe weather conditions.

As for the semantic analysis of stance MWUs, investigation was carried out into the expression *shall be*, since it possesses distinctive meaning in the COMAIR, and thus displays functional variation from its use in general English register. Below are some concordance lines extracted respectively from COMAIR and BNC Baby.

ate power ventilation shall be provided in enclosed cargo  
s. The arrangements shall be such as to provide for at le  
o Operations Briefing shall be used as an opportunity to  
rying timber cargoes shall be selected where possible i  
ations 1997 states: 'It shall be the duty of every worker a  
ety risk assessment shall be used to satisfy the obligat  
adequate handholds shall be a bulwark ladder, such lac  
date and due regard shall be paid to the standards ado  
All lashing materials shall be accompanied by approval  
no lifting equipment shall be used for lifting persons un  
f repose<sup>15</sup>. The ship shall be kept upright during the loa  
scue boat or lifeboat shall be provided with an approved  
he master of the ship shall be supplied with such reliable  
num stability criteria shall be maintained throughout the

Figure 6.4 Snapshot of the concordance lines of *shall be* in COMAIR

unting in the morning . Yeah . So we shall be out er , you know early in the 'Oh well  
right ? Yeah , erm okay well I 'll go , I shall be coming back here actually , but er , I  
nd me about that tonight then Mm , I shall be out tonight mm , before you do it thou  
ekend after 's wedding ? Yeah . So I shall be enjoying myself at the wedding Mm .  
n't both . So I said well no um er er I shall be by myself . They say well no we woul  
bring all right . You know I always we shall be too late by the time we get out The sh  
sband has no time either you come I shall be alone Hmm hmm Very well I said wel  
to the tree . It was very exciting but I shall be very careful next time . Tim and Pegs  
othing out of the ordinary . Hoo hoo . Shall be quite glad . Ah . Look at this , look m  
? Yeah please . Ice ? Yes please . I shall be partying tomorrow night . Oh are you  
ening . Mm it 's still extra calories . I shall be drinking some tonight no doubt and a  
or shall I , yeah I 'll back in , I 'll be I shall be awkward I shall reverse in . Where 's  
at ten o'clock by Friday I du n no who shall be the worst Yes . for the third of Januar  
 . Yes , I 'll be better at ten o'clock , I shall be better at ten o'clock by Friday I du n  
What in here ? Everywhere I go . Oh I shall be quiet . It 's all anonymous , you 're no  
absolute definition of the kind of text I shall be examining . I make certain assumptio  
persons , both or all of those persons shall be treated as jointly and severally liable

Figure 6.5 Snapshot of the concordance lines of *shall be* in BNC Baby

As can be seen from the above concordance lines, there are various subjects for *shall be* in the COMAIR, ranging from criteria to lifeboats. Clearly, the functional meaning of *shall be* in these examples involves command, obligation and determination about something that certainly will or must happen. However, in general English register, where *shall be* usually co-occurs with the subject *I* or *we*, the use of such expression is simply to indicate futurity. It thus can be said that the difference in the use of *shall be* across two corpora best reflects the semantic feature of stance MWUs in the MAIR genre.

## 6.4. Discourse organizing MWUs

This section is intended to discuss the prominent features of discourse organizing MWUs in MAIR genre. To achieve this, the notion of discourse organizing MWUs is first introduced. What follows is an in-depth analysis of these MWUs from functional and semantic perspectives.

### 6.4.1. Notion of discourse organizing MWUs

According to Biber *et al.* (1999), discourse organizing MWUs serve two major functions: topic introduction/focus and topic elaboration/clarification. As the names imply, the first category of MWUs signals that a new topic is coming up; While the topic elaboration/clarification MWUs are used to provide additional explanation or clarification of the topic.

### 6.4.2. Discourse organizing MWUs in COMAIR

As shown in Table 6.1., there are altogether 164 MWUs functioning as discourse organizers in the COMAIR, accounting for approximately 8% of all domain-specific MWUs in terms of types and tokens. Table 6.5 lists some of the MWUs of this type.

**Table 6.5 A list of discourse organizing MWUs arranged in descending order (frequency $\geq$ 40)**

Number	Discourse organizing MWUs	LL value	Frequency
1	in addition	2741.973	512
2	resulted in	2602.831	494
3	as a result	3618.65	483
4	based on	1913.876	405
5	stated that	2199.441	394
6	associated with	2385.155	361
7	found to	668.0307	317
8	continued to	1005.46	313
9	identified that	1023.897	282
10	reported to	677.3814	280

.....

160	was attached to	128.5603	40
161	was considered to be	211.4503	40
162	not considered to be	244.9058	40
163	demonstrates that	205.1936	40
164	informed that	84.8427	40

Further observation within this group also found that the discourse organizing MWUs are used for both topic introduction as well as topic clarification, with almost the same percentage. Table 6.6 below displays the frequency distributions of these two sub-functions.

**Table 6.6 Frequency distributions across subcategories of discourse organizing function (frequency≥40)**

Subcategories	types	%	tokens	%
Topic elaboration	75	45.73%	9204	49.12%
Topic introduction	89	54.27%	9532	50.88%
<b>Total</b>	<b>164</b>	<b>100%</b>	<b>18736</b>	<b>100%</b>

#### 6.4.2. 1. Topic introduction MWUs in COMAIR

Semantically, each sub-function can be distinguished into several types depending on their specific contribution to the discourse. Take the topic introduction MWU as an example, the topics that these MWUs introduce cover not only the factual information of the accidents such as the reasons and results of the occurrence, but also the findings of the investigation, and the recommendations to other vessels. Nevertheless, the MWUs of this sub-function present a noticeable tendency for adopting the pattern of ‘*that*-clause controlled by main verbs in active voice’ to perform the discourse act of topic introduction. Evidence can be found from top 10 most frequent topic introduction MWUs listed in Table 6.7.



Table 6.7 Top 10 most frequent topic introduction MWUs in COMAIR

Number	Topic introduction MWUs	LL value	Frequency
1	stated that	2199.441	394
2	identified that	1023.897	282
3	indicated that	1563.536	276
4	concluded that	1657.976	267
5	indicate that	1418.388	239
6	reported that	1004.678	233
7	confirmed that	1249.821	226
8	recommended that	346.6529	110
9	assumed that	490.8514	94
10	established that	389.7733	92

The heavy reliance on such pattern clearly demonstrates that the reporters give a more prominence to the source of the topics when introducing the topics in MAIR genre. By explicating the information as such, it can place the reporters in an objective position, and more importantly, provide convincing evidence for the following topics. In this respect, divergence emerges when comparing it from general English register. This can be exemplified by the use of '*established that*' in two corpora. Below are the snapshots of the extracted concordance lines.

the loading manual, the investigation has established that Cemford's previous chief officers had  
ended the accident and the investigation established that the accident happened after a failure of the  
previous arrival and departure conditions established that the VCG of the cargo was always left to  
the crew disembarked. Inspection in port established that Good Intent had suffered damage to  
keeping on board, the MAIB investigation established that the company's SMS did not meet the  
ordnance with the COLREGs. It was also established that the bridge manning on Shoreway was  
all four crew. The MAIB investigation<sup>25</sup> established that the collision was caused by a breakdown  
from engine room vents (Figure 2). Having established that there was a fire in the engine room, the  
arg and, in conversation with the master, established that the vessel had grounded earlier in the day  
tory examination of the chart would have established that the master's chosen anchorage location  
on them (grade 8). Atlas (Germany) had established that 12 tie bolts were required to secure the  
ence of external damage. The divers also established that there was no sign of pollution from the  
in the event of flame interruption. It was established that the exhaust temperature of a brand new  
ed out. Tests conducted by Eberspcher established that the temperature of the exhaust gas at the  
ure 7). However, the MAIB investigation established that the warranty packs supplied to other

Figure 6.6 Snapshot of the concordance lines of *established that* in COMAIR



However, it has not yet been established that global warming is due to excessive  
 trees until they had got established. That way you would have a little bit of  
 persisted longest. Having established that the larger skeleton was that of a fi  
 and kinetics Having once established that certain polymeric materials are ca  
 eps there. It was quickly established that the occupier could not be liable un  
 prima facie case had been established that A was liable for procuring breach  
 . In this way it is quickly established that Inversion of B as in 1.9(8) then yie  
 ris on the royal demesne established that no parishioner should have to pay  
 er shorelines, it has been established that for long periods of time in the Pha  
 erful tool, it must now be established that maximum achievable use of  
 more common. It is well established that in the delinquent-prone, home dis  
 existed. Once it has been established that D was committing a criminal offen

Figure 6.7 Snapshot of the concordance lines of *established that* in BNC Baby

As can be seen from the concordance lines in the BNC Baby, the MWU *established that* is mostly used as a passive verb phrase and in 'it BE-tense V-ed that' structure. Clearly, this pattern removes the source of the topics. By doing this, it suppresses the implication of human intervention and put much emphasis on the authority of the knowledge itself.

#### 6.4.2.2. Topic elaboration MWUs in COMAIR

As for the topic elaboration MWUs, they were found to elaborate on the logical relationships rather than providing additional information in MAIR genre. The relationships these MWUs clarify broadly fall into three semantic relation types: causative-resultative relation (e.g., *as a result, it is therefore*, etc.); comparison (e.g., *was contrary to, similar to*, etc.) and adversative relation (e.g., *although not, however although, despite this*, etc.). It has to be noted that these three types are by no means exhaustive, but are believed to be the most common ones within this group. Among these three types, it is not difficult to see that the causative-resultative MWUs take the largest percentage. The wide use of these expressions signals the main conclusions to be drawn from the investigation and highlights the inferences that reporters wants readers to draw from the investigation.

#### 6.4.2.3. Semantic features of discourse organizing MWUs in COMAIR

The semantic features of discourse organizing MWUs can be best reflected in the use of MWUs incorporating with *prompted* and the MWU *associated with*, as these expressions possess distinguished semantic meaning in the MAIR genre. Figure 6.8 and Figure 6.9 display the MWUs incorporating with *prompted* and their concordance lines extracted respectively from COMAIR and BNC Baby respectively.

size been understood, it might have prompted the crew to consider abandoning t  
was approaching, which could have prompted him to take earlier evasive action.  
nce for the VTSO and it should have prompted a high level of attention.V-103 prov  
would have been more likely to have prompted the master to take more interest a  
Primula Seaways, which might have prompted an earlier and more substantial sp  
A.1108(29), it is unlikely to have prompted any change to the boarding arrang  
error with their TMA. This could have prompted them to re-evaluate the situation le  
have been identified. This might have prompted the skipper to avoid the risk and d  
ve been apparent and ought to have prompted a review of his intentions. His app  
result, reference to it might not have prompted any cooling of the third engineer's  
tively. Such a challenge might have prompted the master to review his plan, and  
rt blasts by Alexandra 1 would have prompted them into taking avoiding action at  
marker such as Warning would have prompted quicker action by Ever Smart's bri  
veyors. Such information could have prompted surveyors to pay particular attentio  
CTV coverage of the area might have prompted an earlier response and would hav  
on and Walcon Wizard should have prompted the OOW to use ARPA to determ

Figure 6.8 Snapshot of the MWUs incorporating *prompted* in COMAIR

ewer houses and improved landscaping , prompted 45 letters of objection . They stated the  
Nostalgia combined with a sense of loss prompted Patricia Gosling to visit the store yeste  
leadership in the autumn . The bickering prompted former Tory Premier Sir Edward Heath  
s in as many Premier Division games . It prompted boss Lennie Lawrence to step up his p  
olishers rushed out copies last week and prompted a furore in India . Gower , who averaged  
s school becoming grant maintained has prompted the education secretary to announce n  
attle to gain control of the hotel in 1989 i<sup>a</sup> prompted suggestions that a new attack was imr  
The dire state of the property market has prompted it to slash the book value of its bricks a  
nd the likelihood of drying ground i<sup>a</sup> have prompted Coral 's to cut Docklands Express , my  
res Valenzuela get sucked in , and what prompted his change of heart ? These were the q  
good prizemoney , a situation which has prompted a reduction in Pattern races from 139 t  
backbench Opposition speeches which prompted Mr Christopher Chope , junior environm  
he world 's soft drinks markets has been prompted by the US giants , Pepsi and Coke , w  
r in March this year caused a furore and prompted a special committee of Law Lords to de  
ilaterally by Czechoslovak police . This prompted the Foreign Ministry in Bonn to lodge a

Figure 6.9 Snapshot of the MWUs incorporating *prompted* in BNC Baby

As shown in Figure 6.8, the word *prompted* always combines with *have* to form the

MWU *have prompted* in the COMAIR. By looking at the collocates of *have prompted*, it was found that such an MWU mainly conveys the meaning of ‘cause *sth.* to happen’, and thus bears negative prosody and functions to elaborate a resultative relation in the discourse. Interestingly, modal verbs (e.g., *should*, *would*, *might*, *could*, etc.) always precede the MWU *have prompted* in the COMAIR. This indicates that, when reporters use *have prompted* to bring the results, they tend to express certain stance at the same time. While in BNC Baby, as shown in Figure 6.9, the word *prompted* does not show any preference of co-occurrences. Apart from the meaning of ‘cause *sth.* to happen,’ it also expresses the meaning of ‘encourage/inspire’ in some cases (see 4<sup>th</sup>, 6<sup>th</sup> and 13<sup>th</sup> concordance lines in figure 6.9).

The MWU *associated with* is another typical example which can show significant difference across two corpora. Figures 6.10 and 6.11 well demonstrate this point. In the COMAIR, the MWU *associated with* is normally found in the company of negative items and displays a semantic preference for items from the field of ‘danger’, such as *risk*, *hazard*, etc., as shown in figure 6.10. It thus can be concluded that *associated with* is primed with a negative prosody in the COMAIR. However, this is not the case for general English register. By observing the collocations of *associated with* in the BNC Baby, this MWU was found to co-occur with a variety of words and possess neutral or positive prosody. The change of semantic preference and prosody across two corpora illustrates the distinctive semantic features of discourse organizing MWUs in MAIR genre.



sea anglers did not realise the dangers associated with night fishing so close to a port entrance. Catastrophic failure was the big end bearing associated with the No.6 piston connecting rod. The coroner (section 1.9) have highlighted the risks associated with not completing adequate flooding drills. The damaged stability, such as the risks associated with a compartment fully flooding. However, the Manual that highlighted the dangers associated with CO poisoning from engines (Annex A). The lack of a carbon monoxide alarm. The risks associated with the build-up of CO associated with the risks associated with the build-up of CO associated with the use of canopies. The hazards of canopies also highlighted the extreme hazards associated with shooting pots. Paragraph 14.3.1 of the report would have been fully aware of the risks associated with working on deck. Even if Lee had been vigilant, failures on these occasions were again associated with the propulsion system's hydraulics. It is not discussed tipping doors or the risks associated with repairing dredges. The MCA's 7 Fisheries (15(F), Fishing Vessels: The Hazards Associated with Trawling, including Beam Trawling and paying particular attention to the risks associated with maintenance tasks. Safety recommendations have little available guidance on the risks associated with them. Neither the SeaFish generic risk a

Figure 6.10 Snapshot of the concordance lines of *associated with* in COMAIR

the changing of the clocks is forever associated with the Arts and adenoids. I know the protection rules. The changes are associated with a £2.4m overdraft. Young has agreed, and I first knew him when he was associated with Elspeth Cochrane's literary compares with the 35 fatalities a year associated with angling, 39 with boating, and 44 which lacked the energetic input normally associated with Mansell, banned from yesterday's Jacques de Fouchier, is still closely associated with the group as Chairman Emeritus. When, with a blitheness I had never associated with Pike, he stepped two paces back, I was startled, not an emotion one easily associated with Sergeant Henley, and it might seem where crimes of violence could be associated with the playing of such games. Two other vocabulary, but now suddenly associated with a solid form and gusts of fragrance could not help but benefit from being associated with it. Klepner stepped down from the elements from everyone who had been associated with her and might be helpful. Penelope away. It may, she may have been associated with WI, I don't know. Used to Did he know if not all monarchs and has been associated with a particular and often predominant of social reform, especially those associated with Neville Chamberlain as minister of education for some of the problems now associated with government. POLITICAL CULTURE demonstrated that only one species is associated with dispersal, in contrast to several they become available. The fruits associated with such dispersers are conspicuous. Ericaceae seem to be particularly associated with the attraction of pheasants, at least

Figure 6.11 Snapshot of the concordance lines of *associated with* in BNC Baby

## 6.5. Referential MWUs

This section discusses the patterns and meanings of the domain-specific referential MWUs in the COMAIR and characterizes their most prominent features. Similar to the discussion of stance and discourse organizing MWUs, the notion of referential MWUs is presented first. It is then followed by the detailed description of each functional subcategory.

### 6.5.1. Notion of referential MWUs

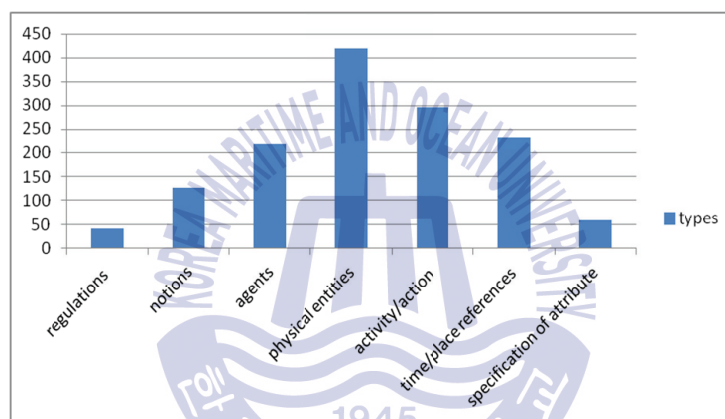
According to Biber *et al.* (2004), referential MWUs ‘identify an entity or single out some particular attribute of an entity as especially important’ (Biber, *et al.* 2004). The MWUs of this type primarily contribute to informational or propositional meanings (Thompson 2000). Within this functional group, there are four major subcategories included: identification/focus, imprecision indicators, specification of attributes, and time/place/text reference, each of which makes their own way to convey the content for the text. As discussed above, the domain-specific MWUs are used to convey the meaning of MAIR genre, thus chiefly function as referential MWUs. Extracted data suggested that the domain-specific referential MWUs in the COMAIR generally make direct reference to physical or abstract entities, notions, actions or processes, space, etc. The semantic categories of these expressions are discussed in the following subsection.

### 6.5.2. Referential MWUs in COMAIR

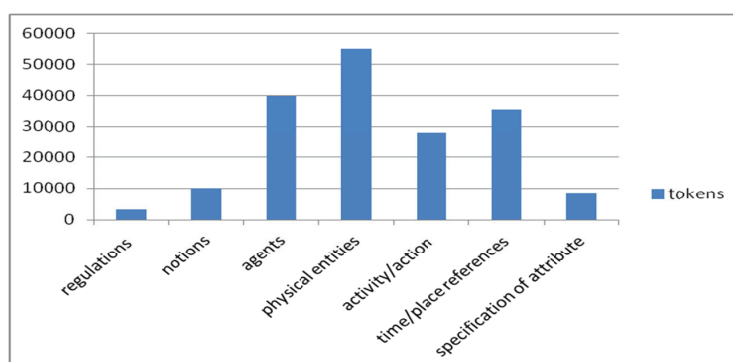
The domain-specific referential MWUs fall into three main categories of the functional classification proposed by Biber *et al.* (2004), namely, a) identification/focus, b) time/place reference and c) specification of attribute. Being the largest category among the three, the ‘identification/focus’ MWUs can be further divided into five semantic groups: the identification of notions, activities/action, regulations, agents (i.e., *organization* and *people*) and physical entities (i.e., *vessels* and *equipment*). Table 6.8 presents the functional classification of the referential MWUs with the corresponding types, tokens and their percentages. Figures 6.12 and 6.13 display the distributions across different functional categories in terms of MWU types and tokens.

**Table 6.8 Frequency distributions across functional subcategories of referential MWUs (frequency $\geq$ 40)**

Functional categories	subcategories	Types	Type (%)	Tokens	Token (%)
<b>Identification/focus</b>	regulations	41	2.94%	3269	1.81%
	notions	126	9.03%	10110	5.60%
	agents	220	15.7%	39903	22.11%
	physical entities	420	30.13%	54975	30.47%
	activity/action	297	21.28%	28164	15.61%
<b>Time/place references</b>		233	16.69%	35680	19.77%
<b>Specification of attribute</b>		59	4.23%	8352	4.63%
<b>Total</b>		1396	100%	180453	100%



**Figure 6.12 Distribution across functional categories of referential MWUs (types)**



**Figure 6.13 Distribution across functional categories of referential MWUs (tokens)**

The above table and figures show that referential MWUs functioning as physical

entity identification form the largest subcategory in terms of types and tokens while the MWUs serving for identifying regulations have the least proportions at 2.94% (types) and 1.81% (tokens). Among all the subcategories, the MWUs identifying activity/action become noticeable, since this group ranks the 2<sup>nd</sup> inter terms of types, but it only has the 4<sup>th</sup> largest number of total occurrences (tokens). The larger types but fewer tokens of this type of MWUs indicate that the use displays a fairly low repetition.

In fact, if token/type ratio is calculated for the repetition rate of each subcategory, it is found that the highest repetition goes to the MWUs identifying agents, with each type occurring 181 times on average. It is then followed by time/place MWUs, which are repeatedly used more than 150 times on average in the COMAIR. The repetition rates of the rest functional categories are relatively low, less than 150 times for the occurrences of each MWU type. Table 6.4 tabulates the repetition rate of each category based on the token/type ratio, which is sorted in a descending order.

**Table 6.9 Repetition rate of functional category based on the token/type ratio**

Rank	Functional Categories	Token/type Ratio
1	Agents	181
2	time/place references	153
3	specification of attribute	141
4	physical entities	130
5	activity/action	95
6	Notions	80
7	Regulations	80

The high repetition of the MWUs identifying agents, time and places demonstrates a low-degree variation of these expressions as the main conveyors of information. It also unveils the most essential elements of MAIR genre, which always include the situational factors such as place and participants. The following subsections describe the salient features of each of the three main functional categories respectively.



### 6.5.2.1. Referential MWUs functioning for identification/focus

Among the domain-specific referential MWUs, a considerable number of expressions fall into the functional category of identification/focus and convey the specific aspects of the content of MAIR genre. To perform such function, this group of MWUs employs a variety of grammatical structures, among which the NP construction has the greatest number. By further examining these NP-based MWUs, it is found that such pattern is primarily used to refer to different types of physical entities. To be specific, NP structure is used either to name the equipment used in maritime domain (i.e., *very high frequency (vhf) radio, the back rope, the steering gear, the bilge alarm*, etc.); to present various types of vessels (e.g., *small fishing vessels, the tanker, the rescue boat*); or to denote different agents (e.g., *the officer of the watch, the master and pilot, the chief engineer*, etc.) and regulations (e.g., *harbor authorities, port state control, coastguard agency*, etc.). As for these NP-based identification MWUs, it should also be noted that most of these expressions have complete NP structure and specialized meanings, thus form as the technical terms in the maritime domain. It is especially the case for MWUs identifying equipment, where almost all the expressions are specialized terminologies.

Unlike NP structure, there exist many differences in the use of other structural types by each subcategory for meaning realization. For example, equipment MWUs employ PP structure to emphasize the equipment as a means of certain activities rather than just a simple entity (e.g., *by VHF radio*). While the use of PP structure by the vessel and agent MWUs is mainly to specify any particular vessels (e.g., *of small fishing vessels, of his vessel*, etc.) and agents (e.g., *by the coastguard, for merchant seamen, from the master*, etc.). In terms of VP structure, it is mostly used by activity/action MWUs to describe the actual activities or actions.

Apart from the above investigation into this group of MWUs, it is of significance to carry out semantic meaning analysis as well. Among these identification MWUs, MWUs identifying vessels, under the subcategory of physical entity, together with activity/action MWUs deserves further attention, since both types display distinctive

features in MAIR genre.

As for vessel MWUs, the majority of these expressions are comprised of the word *vessel*. Even though there are some cases where the MWUs contain *ship*, the similar expressions incorporating *vessel* can also be found in the database, such as all ships<sub>(44)</sub>, all vessels<sub>(140)</sub>, container ship<sub>(108)</sub>, container vessel<sub>(61)</sub>, passenger ship<sub>(53)</sub>, passenger vessel<sub>(113)</sub>, cargo ship<sub>(97)</sub>, cargo vessel<sub>(65)</sub>. Another piece of evidence is obtained from the top 20 MWUs of this type (in Table 6.10 below), in which 15 out of 20 sequences consist of *vessel*. Since *vessel* is known as the formal word expressing a large ship or boat, the extensive use of this word in the MWUs gives an indication of the formal style of the MAIR genre.

**Table 6.10 Most common MWUs identifying vessels in COMAIR**

No.	MWUs	LL value	Freq.
1	the vessel	14077.21	8611
2	the ship	3037.11	2147
3	fishing vessels	10433.49	1418
4	of the vessel	1488.05	1182
5	the boat	1219.47	1020
6	fishing vessel	4080.08	814
7	a vessel	558.89	688
8	that the vessel	1100.01	559
9	to the vessel	76.68	455
10	the liferaft	584.26	401
11	the vessels	88.35	372
12	of the ship	496.06	338
13	vessels in	354.38	336
14	vessels and	86.17	320
15	the vessel to	251.29	297
16	the vessel and	236.97	293
17	the vessel had	490.19	290
18	his vessel	716.44	282
19	the ferry	352.14	281
20	on the vessel	165.68	272

Different from MWUs identifying vessels, the MWUs describing activity/action are of particular interest due to the following two reasons. First, some word sequences are

general in nature, but they convey specialized meaning in the COMAIR. Take the MWUs incorporating *secure* as an example. In the COMAIR, the MWUs combining *secure* are listed in Table 6.11 below.

**Table 6.11 Domain-specific MWUs containing *secure* in COMAIR**

No.	MWUs containing <i>secure</i>	LL value	Frequency
1	secured to	190.73	108
2	was secured	278.68	102
3	and secured	93.87	73
4	secured to the	112.41	70
5	to secure the	158.32	62
6	be secured	199.98	53
7	secured in	72.35	46
8	been secured	152.48	42
<b>Total</b>			556

As shown in the above table, there are 8 different types of MWUs combining *secure*, with the total occurrences of 556 times in the COMAIR. All these expressions adopt VP structure to convey the meaning. By checking the concordance lines of these expressions, it was found that the meaning ‘*secure*’ in these expressions refer to the action of attaching or fastening something firmly, as seen in the following examples from the COMAIR:

- For the passage from Ipswich to Southampton, the towline ***was secured to*** Kingston's towing winch.
- a toolbox talk should have identified that Svitzer Ellerby was unmanned, and so the tug's crew would need to transfer between vessels ***to secure the*** ropes.
- The liferaft ***was secured in*** a cradle on the wheelhouse roof by a Hammar hydrostatic release unit (HRU) and a senhouse slip.

However, by searching the above expressions in BNC Baby, it was found that only few expressions occur in BNC Baby. Despite of their occurrences, they tend to convey different meanings. This can be exemplified by the MWU *to secure the*. In BNC baby, this word sequence mostly express the meaning of ‘protecting or guarding something so that it is safe and difficult to attack, to enter or leave or to obtain.’ Below are some examples extracted from BNC Baby.

N	Concordance
1	. Twisted trails of blame David Ott examines the UN 's tortuous route <b>to secure the</b> prosecution of those committing atrocities in war-torn Bosnia
2	SDLP and Eire politicians who claimed it cost a million pounds each day <b>to secure the</b> border . ☺ The reality is that even if 1,000 soldiers ( as
3	a buyer might be found for the Tyneside yard . Swan was the favourite <b>to secure the</b> lucrative contract until the yard was placed in receivership in
4	patients , and is looking at ways to enable smaller practices to team up <b>to secure the</b> advantages of fund holding . The election result has given the
5	fellow . ☺ Smith and Efford will both have to win if Labour is <b>to secure the</b> 326 seats it needs to form a Government with an overall
6	. The company also backed Sunrise Television ☺ which outbid TV-am <b>to secure the</b> breakfast franchise ☺ where it has a 20 p.c. stake . Scottish
7	, including chairman John Dyer , yesterday lobbied key MPs in an effort <b>to secure the</b> report 's earliest release . An early day motion has also been
8	yesterday as the Barlow Clowes Investors Group ( BCIG ) sought <b>to secure the</b> release of the Parliamentary Ombudsman 's report into the
9	the window . ☺ We 'll do what we can , but let's not forget we are here <b>to secure the</b> ship , and the best place to do that is from here . We ca n't
10	Theresa had fought against even whilst realising there was no other way <b>to secure the</b> loan she needed . And above all it was Mark who had made
11	, but by endorsing it to G he had gone far beyond what was necessary <b>to secure the</b> return of it to A . ( 2 ) Abusing possession Abuse of
12	coalition was especially useful in providing the party with allies more able <b>to secure the</b> cooperation of the working class . Although the Nigeria
13	control over the distribution of resources and professionals seeking <b>to secure the</b> autonomy of their professional role . Yet we can find

**Figure 6.14 Snapshot of the concordance lines of *to secure the* in BNC Baby**

Another salient feature of this group of MWUs is that the description of activities is usually accompanied by nominalization of certain verb constructions, such as *ship handling, for the passage, alteration of course, the list, lifting operation, release of*, etc. In fact, the corpus data suggests that almost half of the MWUs in this group are nominal sequences, which reflect the frequent use of normalization in MAIR genre to express activities. Nominalization is one of the most prominent stylistic features in formal writings, a device which is often used to assign technical legal values to actions or utterances (Vázquez Orta 2010: 273). As Quirk *et al.* (1985:13) points out, “more complex grammatical correlates are to be found in the language of technical and scientific description ... and clauses are often nominalized.” Semantically, nominalization shows different degrees of abstractness of concepts or processes, being characteristic of logical thinking, which is in turn represented by explaining,

reasoning and inferring.

In the COMAIR, a typical example of this phenomenon is the normalized form of the verb *pass*. It was found based on the database that there is a tendency in the COMAIR to use its normalized expressions *passage to*, *passage from*, *passage through* to show the process of *passing* rather than the action of *pass* itself. We randomly selected concordances in the COMAIR.

- By 0919, Doughty had aborted *passage to* the scene and returned to harbour owing to the swell conditions
- The discussions between the master and the assistant pilot focused more on the berthing operation than the *passage to* the berth, and specific roles were not explained.
- This would also have ensured their next day's fishing had continuing south-westerly wind or sea state prevented a *passage from* Torquay on 29 January.
- Less than 5 positions were found charted during the *passage from* the anchorage to the Antwerp pilot station.
- During the *passage through* the Solent there was little communication between the bridge team and the pilots.
- Barfleur's *passage through* the harbour entrance had not been planned or monitored in accordance with navigational best practice.

#### 6.5.2.2. Referential MWUs functioning for specification of attributes

Compared with the referential MWUs, the domain-specific MWUs serving for specifying attributes have a relatively lower proportion (4.23% of MWU types and 4.63% of MWU tokens). Although these MWUs are few in number, this group displays the most distinctive features of MAIR genre, thus deserve further investigation.

Further examination of all specification MWUs, it was found that the tangible

framing MWUs outnumber the intangible framing MWUs. This finding is hardly surprising in that the domain-specific MWUs are more likely to convey the content of the text rather than the abstract characteristics or logical relationships in the text.

Semantically, it is interesting to find that some MWUs adopt the concrete rather than abstract part of semantic meaning. By doing this, these expressions functioning as intangible frames in general English, change to function as tangible frames in the COMAIR. One of the typical examples is the MWU *strength of*. Evidence from BNC Baby shows that although the frequency of *strength of* is low in BNC baby, with only 16 times in total, this expression is usually used in a relatively abstract way, such as to describe somebody's quality of being brave, the power or influence that somebody or an organization has or a strong feeling or opinion. For example,

#### Concordance lines of *strength of* in BNC Baby

- Other drug shares have also fallen, so the Footsie Index is giving a misleading impression as to the underlying *strength of* the market.
- The dividend should rise to 12.3p. reflecting the underlying *strength of* the business.
- The *strength of* the partnership is highlighted for Allan, who also works with a number of other artists.
- Mr Skinner's real concern is not United's loyalties in the event of an Anglo-American war but the bargaining *strength of* the US industry in international negotiations.
- No more was discussed in 1918 and, despite the *strength of* the arguments for further collaboration, they would scarcely have carried the party to a greater commitment.

However, there are 64 occurrences of *strength of* in the COMAIR, all of which are used either to describe the natural force (e.g., *the strength of wind/tidal stream*, etc.) or to refer to the ability of an object or material to carry heavy weight (e.g., *the strength of the rope/steel/lifelines*, etc.).



### Concordance lines of *strength of* in COMAIR

- This abrasion damage would have decreased the tensile *strength of* the rope, and therefore increased the risk of it parting under tension.
- Course alterations intended to regain track were insufficient given the *strength of* the tidal stream setting Commodore Clipper off course.
- The bridge team discussed this and other options and concluded that the vessel would not be able to get alongside with the two available tugs, due to the *strength of* the wind.
- The master overestimated the *strength of* the fire-fighter's lifelines and his ability to manually control their loading in the prevailing conditions.
- Such heat would have also reduced the tensile *strength of* the bolts holding the connecting rod bearing cap in place.

It is also the case for the MWU *effect of*. In the COMAIR, the *effect of* mostly emphasizes the influence of natural force (i.e., *weather; wind, water, etc.*) towards vessels. However, in general English, this expression tends to specify the change or result that *sb./sth.* causes in *sb./sth.* else. See the following examples:

### Concordance lines of *effect of* in BNC Baby

- Although they are extremely unlikely to suffer any lasting *effects of* the infection, it does appear to be the case that young.
- Advertising agencies and banking houses are paying substantial bonuses to their high-earning staff to beat the *effects of* a substantial increase in taxes in the wake of a possible Labour victory on April 9.
- For example, wrote in deeply critical terms of the *effects of* government economic policy on the inner city, and the consequences for policing.
- He was now feeling pleasantly intoxicated from the *effects of* a steady supply of alcohol, which had lifted his flagging spirits.
- The *effects of* humans can go far beyond this, in moving plants away from



their natural range so that they appear native in their new homes.

### **Concordance lines of *effect of* in COMAIR**

- You should consider all aspects of the loading on the vessel, the weight of pots and rope, the catch on deck, the pull of the hauler and the *effects of* wind and tide.
- In the majority of these cases the person in the water was initially responsive and able to help themselves before they rapidly succumbed to the incapacitating *effects of* cold water.
- The sea temperature was 7°C and the air temperature was 8°C (0°C taking into account the *effects of* wind chill).
- The reserves of stability or freeboard remaining may be small to counter any adverse *effects of* sea or wind with consequent danger to crew on deck or to the vessel itself.
- The risk assessment's control measure of using dust masks to limit the *effects of* toxic fumes was flawed, and potentially provided a false sense of security for the crew.

Apart from the above situations, there are also some occasions, where the MWUs serve the same function in both MAIR genre and general English register, but they are used to specify different attributes or convey different semantic meanings across two corpora. One typical example is the MWU *appreciation of*. In BNC Baby, this word sequence usually expresses gratitude or a sympathetic understanding of *sth*. Thus it bears positive prosody in the general English register. Below are the sentences comprising of *appreciation of* in the BNC Baby.

the TEC's intention to improve appreciation of the benefits of training among properly defined and there is an appreciation of what such systems can do and , I feel I must record my great appreciation of the real care and attention I en Scots and Jews arise out of appreciation of herrings . I also came upon a e artist's career i<sup>a</sup> that is , our appreciation of his achievement as a whole re Martin exclaimed in genuine appreciation of the table , beautifully decorate ke . Rufus , who had n't much appreciation of nature usually , nevertheless Barnet . Even Sean's obvious appreciation of her had served mainly to boos to have a much more sensitive appreciation of the social and political d a first response or simply an appreciation of the words , their structure , phenomena . Thus for a proper appreciation of the distinction between " slow actions . These begin from an appreciation of the complementary nature of t a person is doing requires an appreciation of the kind and content of the cess goals as leading to public appreciation of the Engineers' work , but not eckless drivers , and of a wider appreciation of the risks created and the mise writer would like to express his appreciation of the art historical studies of

Figure 6.15 Snapshot of the concordance lines of *appreciation of* in BNC Baby

By contrast, a check of the concordance lines of *appreciation of* in the COMAIR indicated that, despite of the similarity shared with BNC Baby in terms of function, the meaning and semantic preference of this word sequence is deviant from general English. That is, it normally refers to the awareness of a dangerous situation and prefers to collocate with the negative items from the field of 'danger', such as *risk*, *hazard*, *poor*, *lack*, etc. It is thus clear that *appreciation of* bears a negative prosody in the COMAIR. See the following examples:

ding. The master lacked an appreciation of the vessel's likely rate of drift shou was because he lacked an appreciation of Chou Shan's likely rate of turn and ult of the master's lack of appreciation of the dangers resulting from tidal eff ought to have had a better appreciation of the local weather conditions. He ndicated a particularly poor appreciation of the required safety and operating se. Had they done so, their appreciation of the onboard hazards and risks, fety awareness course, his appreciation of the onboard hazards and risks wo s risk assessment and his appreciation of the risk in sailing might have been reness training course, his appreciation of the risk associated with attemptin that the RYA had a proper appreciation of the risks prevalent in the sport. Th

Figure 6.16 Snapshot of the concordance lines of *appreciation of* in COMAIR

Under this category, the MWUs with *heading* as the node word can also demonstrate the use of domain-specific MWUs in the COMAIR deviant from general English. A close scrutiny of the concordance lines of ‘*heading*’ in each corpus revealed that, in BNC Baby, the word is mostly used to combine into a verb phrase ‘*heading for*’, meaning ‘to move in a particular direction’, as can be seen in the following extracted examples.

number of arts publications . TOP folk musicians will be heading for Darlington next month . The Arts Centre is hosting  
 nt , will give a talk on where he considers the world is heading . The event , organised by Durham Friends of the Ea  
 on a couple 's sofa while they slept on j£5,000 offer . Heading the list of dismissals was Middlesbrough manager Ji  
 into claims that the world-famous Mersey Ferries are heading for closure . The call came from trade union leaders i  
 drugs deal went wrong . Det Sgt Greg Cooper , who is heading the murder inquiry , said : j© Clarke is extremely dar  
 f instead of Rosyth in Fife . The Government could be heading for a damaging backbench revolt of Scottish Tories if  
 ciously by their phone for news from Alaska , the man heading the search confirmed last night that it had been susp  
 air coloured hair . The gang told the woman they were heading towards Widnes , and they then set off in a blue car  
 into claims that the world-famous Mersey Ferries are heading for closure . The call came from trade union leaders i  
 sited the Birkenhead headquarters of Cewtec , before heading for Liverpool . At GPT 's Edge Lane headquarters Mr  
 rs grow for ferries FEARS that the Mersey Ferries are heading for closure led to a demand for an inquiry today j- Thi  
 is close to talks with Israel , told Reuters . j© We are heading quickly toward that . The next 24 hours will see a pol  
 years or so he lived the nomadic life of the minstrel , heading off in his car round the well worn folk circuits of Gerr  
 elf a fantastic new job . He and his wife Alexandra are heading off to the sun and fun of Mauritius where he will head  
 Thurrock , said he believed the Government was now heading for a j© crazy course of action j- to privatise services  
 part of his education . j- PLAY ME JULIAN DICKS is heading for a bust-up with manager Billy Bonds if he 's left ou  
 rmand Felten 's outstretched fists a split-second after heading one of Wednesday 's eight goals in their UEFA Cup i

Figure 6.17 Snapshot of the MWUs incorporating *heading* in BNC Baby

While in the COMAIR, there are altogether five different MWUs containing the word *heading* (see Table 6.7). Among these expressions, *heading* tends to be used to form a NP-based pattern. The typical construction is ‘(on) a *heading of*’.

Table 6.12 MWUs combining *heading* in COMAIR

MWUs	LL value	Frequency
a heading of	343.06	77
on a heading of	225.68	48
heading to	19.91	45
heading and	21.14	44
heading was	52.98	44

By further checking the concordance lines of the MWU (*on*) a heading of, it is not difficult to find that the typical collocates of this MWU are the concrete degree of the angle. Hence it can be said that the MWUs with *heading* was more likely to serve as a tangible frame in the COMAIR for specifying the heading angle of the vessels, as seen in the following excerpts:

Between 0147 and 0214, Islay Trader was stationary on a heading between 052° and 063°. The visibility was good and the sea was calm. The second officer replied '174jã±. The pilot then ordered a heading of 157° quickly followed by 150°. During the vessel's turn to port, the pilot set to starboard and, at 1257.26, the pilot ordered a heading of 163°. Ocean Prefect's master and the pilot were discussing the position of the two buoys (Figure 4) and the pilot steadied the vessel on a heading of 167°. The engine telegraph was at 'slow ahead' and the vessel was moving slowly. At 1254, the pilot steadied the vessel on a heading of 167°, the channel's axis, but it was immediately set to the starboard. Ocean Prefect passed between No.3 and No.4 lateral posts on a heading of 163° at 4.8kts. Seconds later, there was an exchange of messages between the two vessels. The pilot immediately ordered a heading of 150° followed by hard-a-port and 'full ahead'. Ocean Prefect's master ordered to bring Huayang Endeavour around slowly on to a heading of about 170° in a series of smaller manoeuvres. No sound signals were just 655m apart when Seafrontier steadied on to a heading of 172° (Figure 3). Unsure of what Seafrontier's master intended, Primula Seaways was 0.97nm. The ferry continued on a heading of 235° (Figure 8) and its port and starboard sidelights were visible. City of Rotterdam passed the Grimsby Middle buoy on a heading of 125° (Figure 4). At that time, the Denmark registered motor vessel was on a bearing of 138°, whereas the vessel was actually on a heading of 105°. In addition, Primula Seaways was on the starboard side. Next 4 minutes, brought CMA CGM Vasco de Gama onto a heading of 260°. During the turn, the vessel passed 0.25nm north of the buoy. CMA CGM Vasco de Gama passed abeam of Cap Hatteras on a heading of 239° in a position 0.5nm to the east of the Prince Consort buoy. At about 30°, brought CMA CGM Vasco de Gama on to a heading of 260° and started his turn into the Thorn Channel 3 cables

Figure 6.18 Snapshot of the concordance lines of a heading of in COMAIR

### 6.5.2.3. Referential MWUs functioning for time/place reference

As for the functional category of time/place reference, analysis was primarily carried out semantically. The result shows that there also exist semantic differences across two corpora, which can be best illustrated by the MWUs with *forward*. In the COMAIR, *forward* expresses technical meaning, referring to the front part of the vessels. By grouping the MWUs with *forward*, it was found that the meaning of *forward* is realized by various patterns. To be specific, there are altogether 14 different types of MWUs containing *forward*, with the total frequency at 1,410 times. These expressions are tabulated in Table 6.13 below.



Table 6.13 MWUs with *forward* in COMAIR

No .	MWUs incorporating <i>forward</i>	LL score	Freq.	Collocate examples (window span to +5)
1	the forward	621.58	551	tanks, console, deck, cabin, bilge suction, bulkhead
2	forward of	74.11	109	the pots, the wheelhouse, the hold, the bow
3	forward and	78.13	109	aft cabin, an open deck aft, aft working decks
4	forward of the	171.25	90	main engine room, bosun's store, bow, cargo hatch,
5	forward end	511.09	66	of the top deck, of the engine, of the keel,
6	forward mooring	455.52	62	deck, party, area, rope, station,
7	in the forward	118.20	60	cabin, store room, end of hold, machinery space,
8	at the forward	201.78	58	End, mooring station, end of the keel, seating area
9	to the forward	46.77	56	Swimming pool, mooring deck, part of the bridge,
10	on the forward	119.04	53	deck, port side, hatch, mooring deck, console,
11	the forward mooring	273.52	52	deck, station, party, lines, ropes, team,
12	the forward end of	150.27	49	the cabin, the keel, the top deck, the winch,
13	of the forward	34.64	49	mooring deck, bulkhead, mooring deck, tanks, hold,
14	forward end of the	94.67	46	hold, working deck, console, mooring station,

As can be seen from these instances and their concordance lines, *forward* is used either as an attributive or predicative adjective to modify the nouns. Furthermore, the collocate examples in Table 6.13 also suggest that the MWUs with *forward* are likely to co-occur with the specific places on board rather than the vessel as a whole (e.g., *forward engine room*, *forward mooring deck*, *forward port side*, etc.). When it comes to refer to a part of the vessel, it tends to be used in the collocation of *forward end* instead of *forward part*.

Rather than carrying the subject-specific meaning, the word *forward* in general English register is inhabited in other type of construction and conveys different meanings. For example, it is usually used as an adverb in a VERB + ADV pattern with a range of meanings, such as (1) towards a place or position that is in front (e.g., *lean forward*, *run forward*) and (2) towards a good result in the future (e.g., *look forward*, *put forward*), as shown in the concordance lines below.

agree with comments made by others that the way forward lies in a package of investment in small business. A cabinet minister and founder member of the SDP, put forward this thesis last week when she flew back to B. It can only do that, however, if more people come forward to share their lives for a day or two each month. The recreation committee, has revealed she is looking forward to seeing the controversial production of Swan. Indeed, a similar interpretation was put forward by President de Klerk yesterday when he said. I agree with comments made by others that the way forward lies in a package of investment in small business. A cabinet minister and founder member of the SDP, put forward this thesis last week when she flew back to B. It can only do that, however, if more people come forward to share their lives for a day or two each month. The recreation committee, has revealed she is looking forward to seeing the controversial production of Swan for it about a month ago and I'm thoroughly looking forward to seeing it. AN ART exhibition on lesbianism. The Government inspector would expect them to put forward a preferred route as a matter of policy. The community. Middlesbrough Council will be putting forward a strong case for a core area, including. The with a complaint should pass them on and we will forward them to the Office of Fair Trading, he said. when seeking consent for an overhead line to put forward proposals in the alternative. It's the fairest way much that there is difficulty in carrying these tests forward. They are very much in the interests of the

Figure 6.19 Snapshot of the MWUs incorporating *forward* in BNC Baby

## 6.6. Summary

In general, this chapter has discussed the semantic and functional meaning of the domain-specific MWUs. The functional interpretation has revealed three primary functions that the domain-specific MWUs perform in the COMAIR and the typical patterns they employ to realize these functions, with the identification/focus and specification MWUs being the focus of the investigation. By comparing with general English register, the semantic analysis of the domain-specific MWUs has provided ample evidence to show the distinctive meanings the target MWUs possess in the COMAIR. All the findings are believed to reflect the most salient features of the domain-specific MWUs in MAIR genre.

## Chapter 7

### Conclusions and Implications

This chapter summarizes the major findings of the current research and discusses implications of these findings. It ends with addressing the limitations and providing the recommendations for future studies.

#### 7.1. Summary of the major findings

The present study is a systematic corpus-based investigation of the domain-specific MWUs in MAIR genre with a view to characterizing their salient patterns and meanings. To achieve this principal objective, the target MWUs were first identified by applying a new approach, which incorporate the notion of ‘meaning’ into statistical-based measure. This newly proposed method ensures the domain-specific MWU extraction to the largest extent and provides valid data for the subsequent analysis. The major findings and their implications are summarized as follows.

First, the domain-specific MWUs are largely composed of 2-word sequences, while the occurrences of 4- and 5-word MWUs are relatively rare in the COMAIR. Among all the target MWUs, only 1.10% of the expressions occur very commonly within the genre ( $>1,000$  times). However, the majority of the expressions (70.97%) occur with the frequency less than 100 times. The skewed distribution indicates that MAIR genre tends to employ a wide variety of domain-specific MWUs rather than repetition of a small number of common expressions. One possible reason is that the content of each MAIR is of its nature so diverse and extensive. It may also be due to the ESP-based nature of this genre, which brings about many technical terms and idiosyncratic expressions with low frequency.

In terms of the syntactic features of the domain-specific MWUs, NP-based structure is the most commonly employed grammatical type, accounting for approximately 60% of MWU types and tokens. The abundant use of this structure



implies that the domain-specific meaning of MAIR genre is largely carried in the nominal group. The meanings that these word sequences convey cover a broad range in maritime domain including notions, names of entities, regulations, agents, and so on. They all provide ample evidence for understanding the meaning construction of the domain-specific MWUs in MAIR genre. Apart from NP structure, there is also a marked prevalence of VP structures among the domain-specific MWUs in the COMAIR and these MWUs present structural variation. Of all the VP-based patterns, the 'verb phrase with active verb' pattern stands out since it incorporates a large number of action verbs, which are used to describe the actions done by the people. The wide use of these phrases implies that MAIR genre tends to highlight the people's roles during the accidents, with particular attention to the information about what or who caused or performed the activity. This finding reflects the characteristics of MAIR genre, which is distinguished from other formal writings. In the COMAIR, PP structures are also commonly employed by the domain-specific MWUs for meaning realization, especially those beginning with *of*, which were used to specify possessions, such as parts of the vessels; accidents, people, etc. It can be inferred that the information that provided in MAIR genre tends to be concrete and specific.

For the meaning analysis of the domain-specific MWUs, it includes both functional and semantic aspects. In general, there are a small number of domain-specific MWUs functioning to express stance in the COMAIR, accounting for about 14% of all target MWUs. Within this group, most MWUs express impersonal epistemic stance, with the purpose of minimizing the imposition of the reporters' opinions. This reflects the stance feature of MAIR genre, where objectivity is essentially required. When realizing epistemic function, these expressions tend to employ the pattern of 'copula *be* + adj.' to express the degrees of both certainty and uncertainty. Other word sequences appear to be deontic in nature, as they are mainly realized by the MWUs incorporating with *require* or modal verbs. The primary function of these MWUs is to set out the obligations and issue suggestions for the agents according to certain norms and regulations. It thus can be argued that the obligation MWUs often shunt back and forth between real world events and laws and tend to establish a relation between

these two. By disclosing the source of norms, it enables the reporters to take an objective stance.

Compared with the other two functions, the domain-specific MWUs functioning as discourse organizers have the least number of both types (8.98%) and tokens (8.18%). Despite of few occurrences, this group displays distinguished features. For the MWUs functioning as topic introduction, they present a noticeable tendency for adopting the pattern of '*that*-clause controlled by main verbs in active voice' to perform the discourse act. The heavy reliance on such pattern clearly demonstrates that the reporters give a more prominence to the source of the topics when introducing the topics in MAIR genre. By contrast, when domain-specific MWUs are used for elaborating the topics, they tend to clarify the logical relationships rather than providing additional information in MAIR genre. Of all kinds of relations, the causative-resultative relation takes the largest percentage.

The referential MWUs are among the commonest types in the COMAIR, making up the dominant proportion at around 77%. This result demonstrates that the primary function of domain-specific MWUs is to express referential meanings and contribute to the thematic development of MAIR genre propositionally. Within this group, the MWUs identifying agents, time and places display a fairly high repetition. This finding demonstrates a low-degree variation of these expressions as the main conveyors of information. It also unveils the most essential elements of MAIR genre, which always include the situational factors such as place and participants. Apart from these, the MWUs describing activity/action are of particular interest. This is not only because some word sequences are general in nature, but they convey specialized meaning in the COMAIR. It is also due to reason that the description of activities is usually accompanied by nominalization of certain verb constructions. Under the specification category, the tangible framing MWUs outnumber the intangible framing MWUs, which demonstrates that the domain-specific MWUs are more likely to convey the content of the text rather than the abstract characteristics or logical relationships in the text. More strikingly, the use of some MWUs is significantly deviant from general English. To be specific, some MWUs adopt the concrete rather

than abstract part of semantic meaning in the COMAIR, which results in the change of function. There are also some occasions, where the MWUs serve the same function in both MAIR genre and general English register, but they are used to specify different attributes or convey different semantic meanings across two corpora. This different semantic preference presented in the COMAIR characterizes the semantic features of MAIR genre.

In all, by gaining insights into the salient features of the domain-specific MWUs in the COMAIR, it is believed to provide valuable understanding of MAIR genre.

## **7.2. Implications of this study**

The present study may yield certain practical implications at three levels.

(1) For domain-specific MWU extraction, the present study has proposed a new approach combining the notion of keyness into statistical-based measure. It has been proved to identify the domain-specific MWUs with greater precision, thus, improve the validity and effectiveness of phraseological studies to a certain extent. In this regard, the newly proposed approach may have potentially useful applications in other ESP discourse.

(2) The present study has also made contribution to the compilation of maritime English corpus. Maritime English corpus, as one of the essential language database in maritime domain, provides authentic English expressions for experts and learners specialized in this field. To fully represent the linguistic features of maritime domain, the corpus usually comprises a wide range of sources. In the present study, the approximately two million self-built specialized COMAIR consists of marine accident investigation reports, which are widely regarded as an essential subgenre in maritime domain. In this respect, it is believed that the COMAIR can be used as a subset of the maritime English corpus.

(3) Pedagogically, this study has described the domain-specific MWUs that may not be detectable by personal intuition. More importantly, it has gained insights into the formulaic nature of MAIR genre with a set of distinctive features, all of which can

serve as a starting point for learning and teaching practice. Through identifying a large quantity of domain-specific MWUs from the COMAIR, the findings are especially useful for the teaching and learning of maritime-specific MWUs. Moreover, the findings may also contribute to providing reference for writing MAIR for the maritime experts who are from non-native English speaking countries,

### **7.3. Limitations of this study**

As an initial attempt to investigate the salient patterns of domain-specific MWUs in MAIR genre, this study suffers from some limitations.

The present study mainly focuses on exploring the contiguous domain-specific MWUs in the COMAIR, but the features that discontinuous MWUs display has not touched upon. This is partly due to the unsolved technical issues for extracting discontinuous MWUs. However, the characterization of the overall patterns in MAIR genre needs to include the analysis of discontinuous MWUs as well. This thus lies in the purpose of future studies.

Another limitation of the current research is that the present investigation has been only carried out from a monolingual perspective. But bilingual context, as a complement to monolingual analysis, helps extend phraseological information across languages and provide a wider range of insights into phraseological features of MAIR genre. For this reason, a future comparative study may be conducted between English speaking countries and other non-native English speaking counterparts.

## References

- Aijmer, K., 1996. *Conversational Routines in English: Convention and Creativity*. London & New York: Longman.
- Aisenstadt, E., 1981. Restricted Collocations in English Lexicology and Lexicography. *ITL Review of Applied Linguistics*. 53, pp.53-61.
- Alexander, R. J., 1978. Fixed expressions in English: A Linguistic, Psycholinguistic, Sociolinguistic and Didactic Study (Part I). *Anglistik und Englischunterricht*, 6, pp.171-188.
- Altenberg, B. & Eeg-Olofsson. M., 1990. Phraseology in spoken English: Presentation of a project. In J. Aarts & W. Meijs. Eds., *Theory and practice in corpus linguistics*. Amsterdam & Atlantic: Rodopi, pp.1-22.
- Altenberg, B., 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations. In Cowie, A. P. eds., *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press, pp.101-122.
- Amosova, N.N., 1963. *Osnovui anglijskoy Frazeologii*. Leningrad: University Press.
- Arnold, I.V., 1986. *The English Word*. Moscow: Vuisshaya Shkola.
- Baker, P., Hardie, A., & McEnery, T., 2006. *A Glossary of Corpus Linguistics*, Edinburgh: Edinburgh University Press.
- Barnbrook, G., 1996. *Language and Computers. A Practical Introduction to the Computer Analysis of Language*. Edinburgh: Edinburgh University Press.
- Biber, D., 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, D., Conrad, S. & Reppen, R., 1998. *Corpus Linguistics: Investigating Structure and Use*. Cambridge University Press, Cambridge.
- Biber, D., Johansson, S., S., Leech, G., Conrad, S., & Finegan, E., 1999. *Longman Grammar of Spoken and Written English*. London: Longman.
- Biber, D., Conrad, S. & Cortes, V., 2003. Lexical bundles in speech and writing: an initial taxonomy. In Wilson, A., P. Rayson, & T. Mc Enery. eds., *Corpus*

- Linguistics by the Lune: A Festschrift for Geoffrey Leech*, Peter Lang, Frankfurt, pp.71-92.
- Biber, D., Conrad, S. & Cortes, V., 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), pp.371-405.
- Biber, D., 2006. *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam/ Philadelphia: John Benjamins Publishing Company.
- Biber, D. & Barbieri, F., 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26, pp. 263-286.
- Biber, D. & Conrad, S., 2009. *Register, Genre and Style*. Cambridge: Cambridge University Press.
- Bolinger, D., 1976. Meaning and memory. *Forum Linguisticum*, 1, pp.1-14.
- Bondi, M., 2010. Perspectives on keywords and keyness: An introduction. In: Bondi, M & Scott, M., eds. *Keyness in texts*. Amsterdam: John Benjamins.
- Butler, C. S., 1997. Repeated word combinations in spoken and written text: Some implications for Functional Grammar. In Butler C. S., Connolly J. H., Gattardo R. A. & Vismans R. M. eds., *A Fund of Ideas: Recent developments in Functional Grammar*, Amsterdam: IFOTT, University of Amsterdam, pp.60-77.
- Chapman, S. & Routledge, P., 2005. *Key Thinkers in Linguistics and the Philosophy of Language[C]*. Edinburgh: Edinburgh University Press.
- Chen, Y. H. & Baker, P., 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14, 30-49.
- Church, K. & Hanks, P., 1989. Word association norms, mutual information, and lexicography, in *Proceedings of the Twenty-Seventh Annual Meeting of the Association for Computational Linguistics*. Vancouver: University of British Columbia, pp. 76–83.
- Church K. W. & Hanks P., 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16 (1), pp.22-29.
- Church, K., Gale, W., Hanks, P. & Hindle, D., 1991. Using Statistics in Lexical Analysis. In: Zermick. U., ed, *Lexical Acquisition*. Englewood Cliffs, NJ. Lawrence Erlbaum, pp.115-165.



- Church, K. W. & Mercer, R. L., 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational linguistics*, 19(1),1-24.
- Clear, J., 1993. From Firth principles. Computational tools for the study of collocation. In Baker, M., G. Francis & Tognini-Bonelli E. Eds., *Text and Technology. In honour of John Sinclair*, Philadelphia: Benjamins, pp.271-292.
- Cortes, V., 2002. Lexical bundles in Freshman composition. In Reppen, R., Fitzmaurice S. M. & Biber D. Eds., *Using Corpora to Explore Linguistic Variation*, Amsterdam/ Philadelphia: John Benjamins, pp.131-145.
- Cortes, V., Jones, J. & Stoller, F., 2002. Lexical bundles in ESP reading and writing. Paper presented at *TESOL conference*, Salt Lake City, April 2002.
- Cortes, V., 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23(4), pp. 397-423.
- Coulmas, F., 1981. Poison on your soul: thanks and apologies contrastively viewed. In Coulmas F., Eds., *Conversational Routine*. The Hague, Netherlands: Mouton, pp.69-91.
- Coulmas, F., 1994. Writing systems and literacy: the alphabetic myth revisited. In Verhoeven, L. Eds., *Functional Literacy*, Amsterdam/Philadelphia: John Benjamins Publishing Company, pp.305-320.
- Coulthard, M., 2004. *Advances in Written Text Analysis*. London: Routledge.
- Coulthard, M. & Johnson, A., 2007. *An Introduction to Forensic Linguistics: Language in Evidence*. London: Routledge.
- Cowie, A. P., 1981. The treatment of collocations and idioms in earners' dictionaries. *Applied Linguistics*, 2/3, pp.223-235.
- Cowie, A. P., 1988. Stable and creative aspects of vocabulary use. In R. Carter & M. McCarthy. Eds., *Vocabulary and language teaching*, London, New York: Longman, pp.126-139.
- Cowie, A. P., 1998. *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press.

- Croft, W., 2001. *Radical Construction Grammar*. Oxford: Oxford University Press.
- Daille, B., 1995. Study and Implementation of Combined Techniques from Automatic Extraction of Terminology. In Judith, K. & Resnik P. eds., *The Balancing Act-combining symbolic and statistical approaches to language*. Daudaravičius and Murcinkevičienė 2004, pp. 325-326.
- De Cock, S., Granger, S., Leech, G. & McEnery, T., 1998. An automated approach to the phrasicon of EFL learners. In Granger, S. Eds. *Learner English on Computer*, London and New York: Addison Wesley Longman, pp.67-79.
- De Cock, S., 2000. Repetitive phrasal chunkiness and advanced EFL speech and writing. In Mair C. & Hundt M. Eds., *Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20)*, Freiburg im Breisgau. Amsterdam: Rodopi, pp.51-68.
- De Cock, S., 2003. *Recurrent Sequences of Words in Native Speaker and Advanced Learner Spoken and Written English*. Ph D dissertation, Université catholique de Louvain.
- Dunning, T., 1993. Accurate Methods for the Statistic of Surprise and Coincidence. *Association for Computational Linguistics*, 19(1), pp.61-76.
- Eeg-Olofsson, M. & Altenberg, B., 1996. Recurrent word combinations in the London-Lund Corpus: Coverage and use for word-class tagging. In Percy, C. E., Meyer, C. F. & I. Lancashire. Eds., *Synchronic corpus linguistics: Papers from the sixteenth international conference on English language research on computerized corpora (ICAME 16)*, Amsterdam & Atlanta: Rodopi, pp. 97-107.
- Ellis, N. C., 1996. Sequencing in SLA: phonological memory, chunking and points of order. *Studies in Second Language Acquisition*, 18, pp.91-126.
- Ellis, R., 2008. Phraseology: The periphery and the heart of language. In Meunier, F. & Granger, S. *Phraseology in Foreign Language Learning and Teaching*, Amsterdam/Philadelphia: John Benjamins Publishing Company, pp.1-13.
- Enguehard, C., 1993. Acquisition de Terminologie à partir de Gros Corpus. In *Proceedings of Informatique & Langue Naturelle*, pp. 373-384.

- Erman, B. & Warren. B., 2000. The idiom principle and the open choice principle. *Text*, 20(1), pp.29-62.
- Francis, G., 1995. Corpus-driven grammar and its relevance to the learning of English in a cross-cultural situation. In Pakir, A. eds., *English in Education: Multicultural perspectives*. Singapore: Unipress.
- Francis, G., Hunston, S. & Manning, E., 1998. Collins COBUILD Grammar Patterns 2: Nouns and Adjectives. London: HarperCollins.
- Fraser, B., 1976. The Verb-Particle Combination in English. New York, NY: Academic Press.
- Gerbis, A., 2010. Key words and key phrases in a corpus of travel writing: From early modern English to contemporary “blogs.” In: Bondi and Scott. eds., pp. 147–168.
- Glaser, R., 1998. The stylistic potential of phraseological units in the light of genre analysis. In Cowie A.P. Eds., *Phraseology: Theory, Analysis, and Applications*, Oxford: Oxford University Press, pp.125-143.
- Goldberg, A. E., 1995. A Construction Grammar Approach to Argument Structure. Chicago: The University Of Chicago Press.
- Goldberg, A. E., 2006. Constructions at work. The nature of generalization in language. Oxford: Oxford University Press.
- Gonzales-Rey, I., 2002. La phraseologie du français. Toulouse : Presses Universitaires.
- Granger, S. & Paquot M., 2008. Disentangling the phraseological web. In Granger S. & F. Meunier. Eds., *Phraseology: An interdisciplinary perspective*, Amsterdam/ Philadelphia: John Benjamins Publishing Company, pp.145-160.
- Granger, S. & Meunier, F., 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam/ Philadelphia: John Benjamins Publishing Company.
- Grice, S.T., 2008. Phraseology and linguistic theory: A brief survey. In Granger, S. & Meunier, F. Eds., *Phraseology: An interdisciplinary perspective*, Amsterdam/ Philadelphia: John Benjamins Publishing Company, pp. 3-26.
- Gries, S. T., 2008. Phraseology and linguistic theory: A brief survey, in Granger, S. & Meunier, F. eds, *Phraseology: An Interdisciplinary Perspective*, Amsterdam:

- Benjamins, pp. 3–25.
- Gries, S. T. & Anatol, S., 2010a. Cluster analysis and the identification of collexeme classes. *Experimental and Empirical Methods in the Study of Conceptual Structure, Discourse, and Language*. CSLI: Stanford, pp.73-90.
- Gries, S. T., Beate, H. & Doris, S., 2010b. Converging evidence II: More on the association of verbs and constructions. *Experimental and Empirical Methods in the Study of Conceptual Structure, Discourse, and Language*. CSLI: Stanford, pp.59-72.
- Gries, S. T., 2011. Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions? In: Brdar, M., Gries S. T., and Fuchs M. Ž. eds., *Cognitive linguistics: Convergence and expansion*. John Benjamins: Amsterdam/Philadelphia, pp.237-256.
- Gross, G., 1996. *Lex expressions figées en français: Noms composés et autres locutions*. Paris: Ophrys.
- Halliday, K. A. K., 1993. Quantitative studies and probabilities in grammar. In Hoey, M. Eds., *Data, description, discourse. Papers on the English language in honour of John McH. Sinclair*, London: HarperCollins, pp.1-25.
- Halliday, M, A. K., 1994. An introduction to functional grammar (2nd ed.). London: Arnold.
- Heid, U., 1999. Extracting Terminologically Relevant Collocations from German Technical Texts. <http://www.ims.uni-stuttgart.de/~uli/>.
- Hoey, M., 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
- Hofland, K. & Johansson, S., 1982. *Word Frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities.
- Howarth, P. A., 1998. Phraseology and second language proficiency. *Applied Linguistics*, 19(1), pp.22-44.
- Hudson, J., 1998. Perspectives on Fixedness: Applied and Theoretical [Lund Studies 94]. Lund: Lund University Press.
- Hunston, S., 2002. Pattern grammar, language teaching, and linguistic variation. In

- Reppen, R., M. Fitzmanurice, & Biber, D. Eds., Using corpora to explore linguistic variation, Amsterdam: Benjamins, pp.167-183.
- Hunston, S., 2008. Starting with the small words: Patterns, lexis and semantic sequences. *International journal of corpus linguistics*, 13, pp. 271-295.
- Hunston, S. & Frnacis, G., 1998. Verbs observed: a corpus-driven pedagogic grammar. *Applied Linguistics*, 19, pp.45-72.
- Hunston, S. & Francis, G., 2000. *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam/ Philadelphia: John Benjamins.
- Jhang, S. & Lee, S., 2013a. Collocational networks and semantic preference: A case study of near synonyms of Maritime English vocabulary. *The New Korean Association of English Language and Literature*, pp.31-46.
- Jhang, S. E. & Lee S. M., 2013b. Clusters and key clusters in the Maritime English Corpus. *Journal of Language Sciences*, 20(4): 199-219.
- Jhang, S. E., Kim. S. & Qi, Y. L., 2018. Lexical bundles in ESP writing: Marine accident investigation reports. *Linguistic Research*, 35(Special Edition), pp.105-135.
- Justeson, J., 1993. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. In *IBM Research Report*, RC 18906 (82591) 5/18/93.
- Jurafsky, D. & Martin J. H., 2000. *Speech and Language Processing. An Introduction to Natural Language Process, Computatioinal Linguistics, and Speech Recognition*. Upper Saddle River: Prentice-Hall.
- Jurafsky, D. & Martin J. H., 2009. *Speech and Language Processing*. New Jersey: Pearson Education, Inc.
- Kecskes, I., 1997. A cognitive-pragmatic approach to situation-bound utterances in SLA. Paper presented to *the Chicago Linguistics Society*. March 7, 1997.
- Kecskes, I., 1999. Situation-bound utterances from an interlanguage perspective. In Verschueren, J. Eds., *Pragmatics in 1998: selected papers from the 6th International Pragmatics Conference*, 2. Antwerp: International Pragmatic Association, pp.299-310.

- Keckes, I., 2003. *Situation-Bound Utterances in L1 and L2*. Berlin/New York: Mouton de Gruyter.
- Kjellmer, G., 1991. A mint of phrases. In Aijmer, K. and Altenberg B. eds., *English Corpus Linguistics*, London/New York: Longman, pp. 111-127.
- Krashen, S. & Scarecella, R., 1978. On routines and patterns in language acquisition and performance. *Language Learning*, 28(2), pp. 283-300.
- Knappe, G., 2004. *Idioms and fixed expressions in English language study before 1800: a contribution to English historical phraseology*. New York: Peter Lang.
- Lakoff, G., 1987. *Women, Fire and Dangerous Things*. Chicago: University of Chicago Press.
- Langacker, R. W., 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.
- Lindquist, H. & Levin, M., 2005. Foot and mouth: the phrasal patterns of two frequent nouns. Paper presented at *phraseology 2005*, Louvain-LaNeuve, October 2005.
- Lindquist, Hans., 2009. A corpus study of lexicalized formulaic sequences with preposition + hand. In: Corrigan, B., Edith Moravcsik, E., Ouali, H. & Wheatley, K. eds, *Formulaic Language. Vol. I: Structure, Distribution, Historical Change*, Amsterdam: Benjamins, pp. 239–256.
- Mahlberg, M., 2005. *English General Nouns: A corpus Theoretical Approach*. Amsterdam: Benjamins.
- Mahlberg, M., 2007. Corpus stylistics: Bridging the gap between linguistics and literary studies. In: Hoey, M., Mahlberg, M., Stubbs, M. and Teubert, W. eds., *Text, Discourse and Corpora*, London: Continuum, pp. 219–246.
- Manning D. C. & Schutze H., 2000. *Foundations of Statistical Natural Language Processing*. Massachusetts: The MIT Press.
- Mauranen, A., 2001. Reflexive academic talk: Observations from MICASE. In Simpson, R. C. & Swales J. M. Eds., *Corpus Linguistics in North America: Selections from the 1999 Symposium*, Ann Arbor, MI: University of Michigan Press, pp.165-178.
- Mauranen, A., 2003. “It seems to me you’re saying”: Formulae in academic speech.



- Paper presented at AAAL conference, Arlington, VA., March 2003.
- McEnery, T. & Hardie, A., 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
- Manning, C. & Schütze, H., 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- Mejri, S., 2005. Introduction: polysemie et polylexicalite. In Mejri, S. Eds., *Polysemie et Polylexicalite. Syntaxe et Semantique* ,5, pp.13-30.
- Meunier, F. & Granger, S., 2008. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam/ Philadelphia: John Benjamins Publishing Company.
- Moon, R., 1997. Vocabulary connections: multi-word items in English. In Schmitt, N. & M. McCarthy. Eds., *Vocabulary: Description, Acquisition and Pedagogy*, Cambridge: Cambridge University Press, pp.40-63.
- Moon, R., 1998. *Fixed Expressions and Idioms in English. A Corpus based Approach*. Oxford: Clarendon Press.
- Nation, P., 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.
- Nattinger, J. R., 1980. A lexical phrase grammar for ESL. *TESOL Quarterly*, 14, pp.337-344.
- Nattinger, J., 1988. Some current trends in vocabulary teaching. In R. Carter & M. McCarthy. Eds., *Vocabulary and Language Teaching*. New York: Longman, pp.62-82.
- Nattinger, J. R. & De Carrico, J. S., 1992. *Lexical Phrases and Language Teaching*. Oxford : Oxford University Press.
- Oakes, M., 1998. *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh.
- O'Keeffe, A., McCarthy, M. & Carter, R., 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge University Press.
- Partington, A., 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.
- Partington, A. & Morley, J., 2004. From frequency to ideology: Investigating word

- and cluster/bundle frequency in political debate. In B. Lewandowska-Tomaszczyk. Eds. *Practical Applications in Language and Computer (PALC)*, Peter Lang, pp.179-192.
- Pawley, A. & Syder, F. H., 1983. Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency. In J. Richards & Schmidt, R. eds., *Language and Communication*, London: Longman, pp.191-225.
- Piao, S. L., Rayson, P., Archer, D., Wilson, A., & Mc Enery, T., 2003. Extracting Multiword Expressions with a Semantic Tagger. In *Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, at ACL 2003, 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan, July 12, 2003, pp.49-56.
- Poos, D. & Simpson, R., 2002. Cross-disciplinary comparisons of hedging: Some findings from the Michigan Corpus of Academic Spoken English. In Reppen, R., Fitzmaurice S. M. & Biber D., Eds., *Using Corpora to Explore Linguistic Variation*, Amsterdam/ Philadelphia: John Benjamins, pp.3-23.
- Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J., 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Rayson, P. & Garside, R., 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora, 38<sup>th</sup> Annual Meeting of the Association of Computational Linguistics*. (ACL 2000).Hong Kong, pp1-6.
- Read, T. R. C. & Cressie, N., 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York.
- Renouf, A. & J. Mc H. Sinclair., 1991. Collocational frameworks in English. In Aijmer, K. & B. Altenberg. Eds., *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, London and New York: Longman, pp.128-143.
- Römer, U., 2005. *Progressives, Patterns, Pedagogy: A corpus-driven approach to English Progressive Forms, Functions, Contexts and Didactics*. Amsterdam: Benjamins.
- Römer, U., 2009. The inseparability of lexis and grammar. *Annual Review of Cognitive Linguistics*, 7, pp.141-163.

- Römer, U. & Schulze, R., 2008. Patterns, meaningful units and specialised discourses. *International journal of corpus linguistics* 13(3), special issue. Amsterdam: John Benjamins.
- Römer, U. & Schulze, R., 2009. *Exploring the lexis-grammar interface*. Amsterdam: John Benjamins.
- Santini, M., 2004. A shallow approach to syntactic feature extraction for genre classification. In *proceedings of the 7<sup>th</sup> Annual Colloquium for the UK Special Interest Group for Computational Linguistics*. Birmingham, UK. , pp. 6-7.
- Schmitt, N. & Carter. R., 2004. Formulaic sequences in action: An introduction. In Schmitt. N. eds., pp.1-22.
- Scott, M., 2004. WordSmith Tools version 4. Oxford: Oxford University Press.
- Scott, M. & Tribble, C., 2006. *Textual patterns*. Amsterdam: John Benjamin.
- Scott, M., 2008. *Oxford Wordsmith Tools 5.0*. Lexical Analysis Software: Liverpool.
- Scott, M., 2009. *In Search of a Bad Reference Corpus. What's in Word-list? Investigating Word Frequency and Keyword Extraction*. Ashgate: Oxford.
- Scott, M., 2016. *Wordsmith tools (Version 6.0)*. [Computer Software] Liverpool: Lexical Analysis Software.
- Silva, J. F., & G. P. Lopes., 1999. A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora. In *Proceedings of the VI Meeting on the Mathematics of Language*.
- Simpson, R., 2004. Stylistic features of academic speech: the role of formulaic expressions. In Connor, U. & Upton, T. A. *Discourse in the professions: perspectives from corpus linguistics*. Amsterdam: Benjamins, pp.37-64.
- Simpson-Vlach, R. & Ellis, N. C., 2010. An academic formulas list: new methods in phraseology. *Research Applied Linguistics*, 31(4): 487–512.
- Sinclair, J., 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.
- Sinclair, J., 1991. *Corpus, Concordance and Collocation*. Oxford: Oxford University Press.
- Sinclair, J. 1998. The lexical item. In: Weigand, E. ed., *Contrastive Lexical Semantics*, 128

- Amsterdam: Benjamins, 1–24. Reprinted in Sinclair, J. and Carter, R. eds, *Trust the Text: Language, Corpus and Discourse*. London: Routledge, pp.131–148.
- Sinclair, J., 2001. Review of The Longman Grammar of Spoken and Written English. *International Journal of Corpus Linguistics*, 6(2), pp.339-359.
- Sinclair, J., 2004. *Trust The Text: Lexis, Corpus, Discourse*. London: Routledge.
- Sinclair J., 2004. Progress and Prospects in Corpus Linguistics. *Modern Foreign Languages*, 27 (2), pp.114-128.
- Sinclair, J., 2006. *Linear Unit Grammar: Integrating speech and writing*. Amsterdam/ Philadelphia: John Benjamins Publishing Company.
- Sinclair, J., 2008. The phrase, the whole phrase, and nothing but the phrase. In Granger, S. & Meunier, F. eds., *Phraseology: An interdisciplinary perspective*. Amsterdam/ Philadelphia: John Benjamins Publishing Company, pp.407-410.
- Snedecor, G. W. & Cochran, W. G., 1989. *Statistical Methods*. (8<sup>th</sup> edition) Ames: Iowa State University Press.
- Stubbs, M., 2001. *Words and Phrases. Corpus Studies of Lexical Semantics*. Oxford : Blackwell Publishers.
- Stubbs, M., 2002. Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, 7(2), pp.215-244.
- Stubbs, M., & Barth. I., 2003. Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Language*, 10(1), pp.61-104.
- Stubbs, M., 2005. The most natural thing in the world: Quantitative data on multi-word sequences in English. Paper presented at *Phraseology*, 2005.
- Stubbs, M., 2006. Corpus analysis: The state of the art and three types of unanswered questions. In: Thompson, G. and Hunston, S. eds, *Style and corpus: Exploring connections*. London: Equinox, pp. 15–36.
- Stubbs, M., 2009. Technology and phraseology: With notes on the history of corpus linguistics. In: Römer, U. and Schulze, R. eds, *Exploring the Lexis–Grammar Interface*. Amsterdam: Benjamins, pp.15–31.
- Sugiura, M., 2002. Collocational knowledge of L2 learners of English: A case study of Japanese Learners. In Saito T., Nakasnura J. & Yamazaki S. eds., *English*

- Corpus Linguistics in Japan*. Amsterdam: Rodopi, pp.303-323.
- Svensson, M.H., 2004. *Criteres de figement. L'identification des expressions figees contemporain*. PhD dissertation, Umea University.
- Svensson, M. H., 2008. A very complex criterion of fixedness: Non-compositionality. In Granger, S. & Meunier, F. eds., *Phraseology: An interdisciplinary perspective*, Amsterdam/ Philadelphia: John Benjamins Publishing Company, pp.81-94.
- Swales, J., 1990. *Genre Analysis*. Cambridge: Cambridge University Press.
- Swales, J., 2001. Metatalk in American academic talk: The cases of point and thing. *Journal of English Linguistics*, 29 (1), pp.34-54.
- Teubert, W., 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10(1), pp.1-13.
- Tognini-Bonelli, E., 2001. *Corpus Linguistics at Work*. Amsterddam/ Philadelphia: John Benjamins Publishing Company.
- Thompson, G., 2000. *Introducing Funcitonal Grammar*. Beijing: Foreign Language Teaching and Research Press & London: Edward Arnold (Publishers) Limited.
- Van Lancker., 2004. When novel sentences spoken or heard for the first time in the history of the universe are not enough: toward a dual-process model of language. *International Journal of Language & Communication Disorders*, 39(1), pp.1-44.
- Vázquez-Orta, I., 2010. A contrastive analysis of the use of modal verbs in the expression of epistemic stance in Business Management research articles in English and Spanish. *Ibérica* (19), 77-96.
- Vinogradov, V. V., 1947. Ob osnovnuikh tipakh frazeologicheskikh edinit v russkom yazuike. In Shakhmatov, A. A. eds, *Sbornik statey I materialov*. Moscow: Nauka, pp.339-364.
- Warren, M. & Greaves, C., 2007. Concgramming: a computer-driven approach to learning the phraseology of English. *ReCALL Journal*, 17 (3), pp. 287-306.
- Wei, N. X., 2009. On the phraseology of Chinese learner spoken English: Evidence of lexical chunks from COLSEC. In Jucker, A. H., Schreier D. & M.Hundt. eds., *Language and Computers*, pp.271-296.

- Wray, A., 2000. Formulaic sequences in second language teaching: principles and practice. *Applied Linguistics*, 21 (4), pp.463-489.
- Wray, A., 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.





## Appendix

### The list of domain-specific MWUs used for the present study

No.	Domain-specific MWUs	LL. Value	Freq.	Structural types	Functional types
1	the vessel	14077.21	8611	NP	physical entities (vessel)
2	the master	7963.27	4563	NP	agent
3	on board	23906.63	4119	PP-based fragment	acvitivity/action
4	the skipper	4814.09	2810	NP	agent
5	the crew	2961.44	2625	NP	agent
6	the port	2722.32	2269	NP	place reference
7	the ship	3037.11	2147	NP	physical entities (vessel)
8	the bridge	3176.47	2112	NP	place reference
9	chief officer	18006.34	2060	NP	agent
10	the engine	2491.61	1829	NP	physical entities (equipment)
11	the chief officer	8559.85	1602	NP	agent
12	fishing vessels	10433.49	1418	NP	physical entities (vessel)
13	of the vessel	1488.05	1182	PP-based fragment	physical entities (vessel)
14	engine room	10345.26	1146	NP	place reference
15	would have been	7175.74	1087	other VP-based fragment	stance
16	the pilot	1387.30	1082	NP	agent
17	the starboard	1328.50	1079	NP	place reference
18	the deck	653.68	1022	NP	place reference
19	the boat	1219.47	1020	NP	physical entities (vessel)
20	the wheelhouse	1932.36	1015	NP	place reference
21	ensure that	5289.62	945	THAT CLAUSE	stance
22	the engine room	4941.95	860	NP	place reference
23	the maib	1317.83	815	NP	agent
24	fishing vessel	4080.08	814	NP	physical entities (vessel)
25	chief engineer	6531.40	753	NP	agent
26	in accordance with	4709.43	744	PP-based fragment	specification of attributes
27	fitted with	4284.74	727	VP-P	acvitivity/action
28	to ensure that	3344.91	716	THAT CLAUSE	stance
29	a vessel	558.89	688	NP	physical entities (vessel)
30	to port	850.31	677	TO CLAUSE	place reference
31	on the bridge	2869.59	676	PP-based fragment	place reference
32	the cargo	355.15	661	NP	physical entities (equipment)
33	bridge team	5712.99	652	NP	agent
34	would be	2736.99	646	BE+ADJ./NOUN/PP	stance

35	second officer	5347.16	643	NP	agent
36	vhf radio	6869.02	610	NP	physical entities (equipment)
37	the chief engineer	3123.80	588	NP	agent
38	on deck	1681.45	554	PP-based fragment	place reference
39	the forward	621.58	551	NP	place reference
40	to starboard	1103.00	539	PP-based fragment	place reference
41	the second officer	2791.05	526	NP	agent
42	in addition	2741.97	512	PP-based fragment	discourse organizer
43	the collision	549.51	506	NP	activity/action
44	port side	3556.68	502	NP	place reference
45	the mate	810.92	496	NP	agent
46	resulted in	2602.83	494	VP-A	discourse organizer
47	as a result	3618.65	483	PP-based fragment	discourse organizer
48	unable to	1990.81	481	ADJ. fragment	stance
49	the coastguard	477.15	468	NP	agent
50	starboard side	3874.24	463	NP	place reference
51	operation of	1112.73	460	NP+of fragment	activity/action
52	of the crew	743.30	459	PP-based fragment	agent
53	to the vessel	76.68	455	PP-based fragment	physical entities (vessel)
54	aware of	1624.85	449	ADJ. fragment	stance
55	the master and	568.85	436	other NP fragment	agent
56	main engine	3216.67	418	NP	physical entities (equipment)
57	was fitted	1309.12	418	VP-P	activity/action
58	the harbor	351.72	411	NP	place reference
59	likely that	2101.76	410	THAT CLAUSE	stance
60	the stern	512.50	406	NP	place reference
61	based on	1913.88	405	VP-P	discourse organizer
62	the liferaft	584.26	401	NP	physical entities (vessel)
63	the bridge team	2104.98	401	NP	agent
64	result of	1323.03	397	NP+of fragment	specification of attributes
65	the hull	407.05	396	NP	place reference
66	stated that	2199.44	394	THAT CLAUSE	discourse organizer
67	the passage	218.87	386	NP	activity/action
68	speed of	700.44	378	NP+of fragment	specification of attributes
69	merchant shipping	4460.70	378	NP	agent
70	the winch	438.17	377	NP	physical entities (equipment)
71	the vessels	88.35	372	NP	physical entities (vessel)
72	associated with	2385.15	361	VP-P	discourse organizer
73	required by	1313.69	360	VP-P	stance
74	the watch	217.30	357	NP	agent
75	comply with	2525.45	352	VP-A	specification of attributes
76	crew members	2774.99	351	NP	agent
77	intended to	940.01	344	VP-P	stance

78	a result of	1496.62	342	NP+of fragment	specification of attributes
79	the port side	1627.73	342	NP	place reference
80	on the port	1021.73	342	PP-based fragment	place reference
81	the maritime	337.05	334	NP	notion
82	was required	583.34	333	VP-P	stance
83	is likely to	2004.17	333	BE+ADJ./NOUN/PP	stance
84	possible that	1499.49	331	THAT CLAUSE	stance
85	the grounding	450.63	331	NP	activity/action
86	would not	938.80	329	other VP-based fragment	stance
87	on the starboard	1374.68	329	PP-based fragment	place reference
88	port of	129.69	325	NP+of fragment	place reference
89	as a result of	1786.79	322	PP-based fragment	specification of attributes
90	coastguard agency	3730.82	322	NP	agent
91	the starboard side	1722.81	320	NP	place reference
92	passage plan	3035.80	318	NP	notion
93	maritime and coastguard agent	3848.97	318	NP	agent
94	found to	668.03	317	VP-P	discourse organizer
95	continued to	1005.46	313	VP-A	discourse organizer
96	main deck	2160.58	305	NP	place reference
97	attempt to	1129.98	298	VP-A	stance
98	the engineer	47.99	296	NP	agent
99	compliance with	1966.90	295	other NP fragment	specification of attributes
100	the bosun	538.79	295	NP	agent
101	the passengers	176.53	295	NP	agent
102	the vessel and	236.97	293	other NP fragment	physical entities (vessel)
103	the crane	405.28	293	NP	physical entities (equipment)
104	absence of	1197.64	291	NP+of fragment	specification of attributes
105	is possible	1596.80	289	BE+ADJ./NOUN/PP	stance
106	ability to	1052.64	283	other NP fragment	stance
107	fitted to	416.30	283	VP-P	activity/action
108	it is possible	2171.87	282	IT CLAUSE	stance
109	his vessel	716.44	282	NP	physical entities (vessel)
110	identified that	1023.90	282	THAT CLAUSE	discourse organizer
111	the ferry	352.14	281	NP	physical entities (vessel)
112	the maritime and	835.47	280	other NP fragment	notion
113	reported to	677.38	280	VP-P	discourse organizer
114	was fitted with	1467.48	278	VP-P	activity/action
115	indicated that	1563.54	276	THAT CLAUSE	discourse organizer
116	and therefore	587.68	275	ADV fragment	discourse organizer
117	third officer	2308.86	274	NP	agent
118	responsible for	1750.30	272	ADJ. fragment	stance
119	on the vessel	165.68	272	PP-based fragment	physical entities (vessel)
120	in the wheelhouse	930.94	271	PP-based fragment	place reference

121	the skipper and	360.15	271	other NP fragment	agent
122	the importance of	842.94	270	NP+of fragment	stance
123	the maritime and coastguard agency	3552.76	270	NP	agent
124	was unable to	1161.70	269	BE+ADJ./NOUN/PP	stance
125	the main engine	1231.69	267	NP	physical entities (equipment)
126	concluded that	1657.98	267	THAT CLAUSE	discourse organizer
127	the main deck	1254.01	264	NP	place reference
128	it would have	1474.98	263	IT CLAUSE	stance
129	the hatch	256.93	263	NP	place reference
130	in addition to	1136.19	263	PP-based fragment	discourse organizer
131	as required	873.02	262	ADV fragment	stance
132	the absence of	797.65	258	NP+of fragment	specification of attributes
133	the alarm	89.69	258	NP	physical entities (equipment)
134	to maintain	933.22	256	TO CLAUSE	activity/action
135	other vessels	1202.03	254	NP	physical entities (vessel)
136	to the master	51.97	253	PP-based fragment	agent
137	is possible that	1711.43	252	BE+ADJ./NOUN/PP	stance
138	possible to	485.15	252	ADJ. fragment	stance
139	effects of	936.77	251	NP+of fragment	specification of attributes
140	the chart	211.85	251	NP	physical entities (equipment)
141	relating to	1033.31	251	VP-A	discourse organizer
142	the radar	155.97	250	NP	physical entities (equipment)
143	crew of	14.67	250	NP+of fragment	agent
144	the berth	200.64	249	NP	place reference
145	the lifeboat	235.59	249	NP	physical entities (vessel)
146	control measures	2278.67	247	NP	notion
147	caused by	1456.78	247	VP-P	discourse organizer
148	to enable	970.04	245	TO CLAUSE	stance
149	capable of	1003.90	243	ADJ. fragment	stance
150	ism code	2468.61	241	NP	regulation
151	bow thruster	2883.70	241	NP	physical entities (equipment)
152	effect of	684.34	240	NP+of fragment	specification of attributes
153	cause of	720.63	240	NP+of fragment	specification of attributes
154	it is likely	1735.96	239	IT CLAUSE	stance
155	the port of	319.71	239	NP+of fragment	place reference
156	indicate that	1418.39	239	THAT CLAUSE	discourse organizer
157	it is possible that	1651.60	238	IT CLAUSE	stance
158	to the bridge	227.43	237	PP-based fragment	place reference
159	the bilge	193.55	235	NP	physical entities (equipment)
160	course to	327.55	235	other NP fragment	activity/action
161	reported that	1004.68	233	THAT CLAUSE	discourse organizer
162	the propeller	261.67	231	NP	physical entities (equipment)
163	referred to	941.47	231	VP-P	discourse organizer

164	attached to	828.55	231	VP-P	discourse organizer
165	of collision	406.96	230	PP-based fragment	acvity/action
166	fish hold	2401.89	228	NP	place reference
167	the tank	137.01	228	NP	place reference
168	a ship	210.73	228	NP	physical entities (vessel)
169	considered to	424.47	228	VP-P	discourse organizer
170	was required to	756.87	227	VP-P	stance
171	indicates that	1381.48	227	THAT CLAUSE	discourse organizer
172	the pier	254.77	226	NP	place reference
173	confirmed that	1249.82	226	THAT CLAUSE	discourse organizer
174	to comply with	1204.61	225	TO CLAUSE	specification of attributes
175	were required	633.74	223	VP-P	stance
176	in the event of	1037.72	221	PP-based fragment	specification of attributes
177	the rescue	93.92	220	NP	acvity/action
178	the fishing vessel	671.41	219	NP	physical entities (vessel)
179	the steering	192.97	219	NP	physical entities (equipment)
180	found to be	1556.15	219	VP-P	discourse organizer
181	tidal stream	2789.60	219	NP	physical entities (others)
182	is likely that	1375.48	218	THAT CLAUSE	stance
183	connected to	746.02	218	VP-P	discourse organizer
184	code of practice	2424.87	217	NP	regulation
185	small fishing	1603.08	217	NP	physical entities (vessel)
186	the effect	190.14	217	NP	notion
187	appeared to	815.98	214	VP-A	stance
188	the mooring	103.47	214	NP	physical entities (equipment)
189	the course	13.88	213	NP	acvity/action
190	the vicinity	416.34	212	NP	place reference
191	risk of collision	2083.56	211	NP	notion
192	found that	749.72	210	THAT CLAUSE	discourse organizer
193	the third officer	380.35	210	NP	agent
194	aware that	961.41	208	THAT CLAUSE	stance
195	it is likely that	1449.89	208	IT CLAUSE	stance
196	the seabed	394.44	207	NP	place reference
197	equipped with	1399.24	207	VP-P	acvity/action
198	attempted to	807.69	206	VP-A	stance
199	of the bridge	168.45	206	PP-based fragment	place reference
200	of the port	150.29	206	PP-based fragment	place reference
201	by the master	433.46	206	PP-based fragment	agent
202	the towing	158.96	205	NP	physical entities (equipment)
203	certificate of competency	2630.92	204	NP	notion
204	would not have	1531.68	202	other VP-based fragment	stance
205	the bridge and	271.54	202	other NP fragment	place reference
206	its vessels	649.03	202	NP	physical entities (vessel)

207	to monitor	668.74	202	TO CLAUSE	acvity/action
208	in the vicinity	1082.92	201	PP-based fragment	place reference
209	of navigation	355.37	201	PP-based fragment	notion
210	the hold	54.60	200	NP	place reference
211	the helm	284.38	200	NP	physical entities (equipment)
212	passage planning	2015.80	200	NP	notion
213	the anchor	130.66	199	NP	physical entities (equipment)
214	with the vessel	203.93	198	PP-based fragment	physical entities (vessel)
215	and the master	17.06	198	other NP fragment	agent
216	required to be	990.04	197	VP-P	stance
217	in the vessel	16.12	197	PP-based fragment	physical entities (vessel)
218	the chain	253.24	196	NP	place reference
219	the quay	258.19	195	NP	place reference
220	resulting in	876.67	195	VP-A	discourse organizer
221	noted that	1089.35	194	THAT CLAUSE	discourse organizer
222	to the port	119.51	193	PP-based fragment	place reference
223	rescue boat	1575.60	193	NP	physical entities (vessel)
224	of cargo	101.17	193	PP-based fragment	physical entities (equipment)
225	in port	104.08	191	PP-based fragment	place reference
226	because of	335.45	191	PP-based fragment	discourse organizer
227	skipper of	25.77	191	NP+of fragment	agent
228	of the engine	170.59	190	PP-based fragment	physical entities (equipment)
229	chart plotter	2346.76	189	NP	physical entities (equipment)
230	as soon as	1177.34	189	ADV fragment	discourse organizer
231	on the deck	626.42	188	PP-based fragment	place reference
232	the effects	256.28	188	NP	notion
233	fishing industry	1488.93	187	NP	agent
234	crew member	1437.19	187	NP	agent
235	cargo operations	1274.20	187	NP	acvity/action
236	states that	869.44	183	THAT CLAUSE	discourse organizer
237	shall be	1120.13	182	BE+ADJ./NOUN/PP	stance
238	port and starboard	1705.15	182	NP	place reference
239	the fish hold	1202.53	181	NP	place reference
240	fishing gear	1321.20	181	NP	physical entities (equipment)
241	control system	962.72	181	NP	physical entities (equipment)
242	regard to	718.41	181	ADV fragment	discourse organizer
243	the crewman	151.23	180	NP	agent
244	the deckhand	150.07	180	NP	agent
245	required for	354.61	179	VP-P	stance
246	were required to	752.07	179	VP-P	stance
247	the cabin	155.59	179	NP	place reference
248	a lifejacket	718.05	179	NP	physical entities (equipment)
249	the lift	190.25	178	NP	physical entities (equipment)



250	working practices	1796.23	177	NP	regulation
251	apparent that	901.50	176	THAT CLAUSE	stance
252	length of	364.70	176	NP+of fragment	specification of attributes
253	chief officer and	879.76	176	other NP fragment	agent
254	harbour authority	1707.21	176	NP	agent
255	small fishing vessels	1516.28	175	NP	physical entities (vessel)
256	the effects of	503.44	174	NP+of fragment	specification of attributes
257	for vessels	190.00	174	PP-based fragment	physical entities (vessel)
258	flag state	2111.18	174	NP	agent
259	vicinity of	645.18	173	NP+of fragment	place reference
260	considered to be	1254.18	173	VP-P	discourse organizer
261	the effect of	482.91	172	NP+of fragment	specification of attributes
262	starboard side of	956.23	171	NP+of fragment	place reference
263	showed that	932.91	171	THAT CLAUSE	discourse organizer
264	navigational watch	1608.17	170	NP	agent
265	be fitted	640.69	170	VP-P	acvity/action
266	best practice	1887.05	170	NP	acvity/action
267	the pitch	104.30	169	NP	physical entities (equipment)
268	response to	382.75	169	other NP fragment	discourse organizer
269	second engineer	1268.75	169	NP	agent
270	not possible	720.70	168	ADJ. fragment	stance
271	on board and	504.12	168	PP-based fragment	place reference
272	of the wheelhouse	259.27	168	PP-based fragment	place reference
273	it would have been	998.19	167	IT CLAUSE	stance
274	complied with	1199.19	167	VP-A	specification of attributes
275	scv code	1804.68	167	NP	regulation
276	port side of	793.66	167	NP+of fragment	place reference
277	the crew of	166.09	167	NP+of fragment	agent
278	the barge	226.11	166	NP	physical entities (vessel)
279	the container	151.80	166	NP	physical entities (vessel)
280	safe operation	1234.99	165	NP	stance
281	was aware	520.18	165	BE+ADJ./NOUN/PP	stance
282	a speed of	648.11	165	NP+of fragment	specification of attributes
283	for the vessel	81.06	165	PP-based fragment	physical entities (vessel)
284	the voyage	128.95	165	NP	acvity/action
285	by the crew	431.88	163	PP-based fragment	agent
286	a pilot	204.47	163	NP	agent
287	was found to	591.92	162	VP-P	discourse organizer
288	operating in	364.80	161	VP-A	acvity/action
289	on passage	400.37	161	PP-based fragment	acvity/action
290	the vicinity of	468.58	160	NP+of fragment	place reference
291	the port side of	719.60	160	NP+of fragment	place reference
292	the shore	117.50	160	NP	place reference

293	their vessels	591.59	160	NP	physical entities (vessel)
294	the pump	30.05	160	NP	physical entities (equipment)
295	of the boat	231.53	159	PP-based fragment	physical entities (vessel)
296	with regard to	892.82	159	PP-based fragment	discourse organizer
297	is evident	1063.59	158	BE+ADJ./NOUN/PP	stance
298	in the port	202.66	158	PP-based fragment	place reference
299	of the cargo	330.92	158	PP-based fragment	physical entities (equipment)
300	ensuring that	876.87	157	THAT CLAUSE	stance
301	as required by	1008.83	157	ADV fragment	stance
302	it is evident	1205.90	156	IT CLAUSE	stance
303	necessary to	338.60	156	ADJ. fragment	stance
304	master of	11.20	155	NP+of fragment	agent
305	man overboard	1725.20	154	NP	agent
306	ahead of	250.06	153	ADJ. fragment	place reference
307	board the vessel	581.85	153	VP-A	acvity/action
308	the wheelhouse and	262.76	152	other NP fragment	place reference
309	the port and	111.87	152	other NP fragment	place reference
310	from the vessel	167.91	152	PP-based fragment	physical entities (vessel)
311	the yacht	195.48	152	NP	physical entities (vessel)
312	bridge wing	1394.40	152	NP	physical entities (equipment)
313	unaware of	563.16	151	ADJ. fragment	stance
314	classification society	2144.17	151	NP	agent
315	is required	376.01	150	VP-p	stance
316	fitted on	309.13	150	VP-P	acvity/action
317	on watch	384.99	150	PP-based fragment	acvity/action
318	the fishing vessels	479.72	149	NP	physical entities (vessel)
319	radar display	1589.65	149	NP	physical entities (equipment)
320	to indicate	440.11	149	TO CLAUSE	discourse organizer
321	the skipper of	151.04	149	NP+of fragment	agent
322	when the vessel was	747.64	148	ADV fragment	time reference
323	the likelihood	295.32	148	NP	stance
324	by the vessel	100.06	148	PP-based fragment	physical entities (vessel)
325	water ingress	1391.91	148	NP	physical entities (equipment)
326	marine safety	716.12	148	NP	notion
327	contributed to	596.30	148	VP-P	discourse organizer
328	course of	119.97	148	NP+of fragment	acvity/action
329	likelihood of	593.76	147	NP+of fragment	stance
330	in the vicinity of	670.30	147	PP-based fragment	place reference
331	be considered	642.69	147	VP-P	discourse organizer
332	contrary to	603.49	147	ADJ. fragment	discourse organizer
333	the ism code	222.22	146	NP	regulation
334	pec holder	2071.26	146	NP	place reference
335	of a vessel	279.42	146	PP-based fragment	physical entities (vessel)

336	on the port side	1051.69	145	PP-based fragment	place reference
337	and fishing vessels	694.06	145	other NP fragment	physical entities (vessel)
338	hatch covers	1759.18	144	NP	place reference
339	of fishing vessels	653.75	144	PP-based fragment	physical entities (vessel)
340	the helmsman	243.39	143	NP	agent
341	the surveyor	110.52	143	NP	agent
342	are required	559.25	142	VP-P	stance
343	likely to have	995.05	142	ADJ. fragment	stance
344	on behalf of	725.02	142	PP-based fragment	specification of attributes
345	a port	25.41	142	NP	place reference
346	on this occasion	851.97	142	PP-based fragment	discourse organizer
347	to proceed	523.58	142	TO CLAUSE	acvity/action
348	the passage plan	826.26	141	NP	notion
349	all vessels	455.63	140	NP	physical entities (vessel)
350	a crew	7.45	140	NP	agent
351	the fishing industry	860.31	140	NP	agent
352	the cause of	382.04	139	NP+of fragment	specification of attributes
353	similar to	179.65	139	ADJ. fragment	discourse organizer
354	evident that	727.19	138	THAT CLAUSE	stance
355	the scv code	262.68	138	NP	regulation
356	the back rope	261.67	138	NP	physical entities (equipment)
357	the jetty	144.91	137	NP	place reference
358	a lookout	399.09	137	NP	agent
359	conduct of	328.72	136	NP+of fragment	acvity/action
360	not required	279.58	135	VP-P	stance
361	evidence that	501.29	135	other NP fragment	stance
362	most likely	1267.44	135	ADJ. fragment	stance
363	angle of	417.46	135	NP+of fragment	specification of attributes
364	vessel safety	147.91	135	NP	notion
365	the master of	133.44	135	NP+of fragment	agent
366	carriage of	483.67	135	NP+of fragment	acvity/action
367	the likelihood of	408.14	134	NP+of fragment	stance
368	not possible to	692.08	134	ADJ. fragment	stance
369	engine room and	692.64	134	other NP fragment	place reference
370	cargo hold	948.63	134	NP	place reference
371	a boat	129.30	134	NP	physical entities (vessel)
372	and the skipper	19.01	134	other NP fragment	agent
373	were fitted	409.17	134	VP-P	acvity/action
374	passage to	96.78	134	other NP fragment	acvity/action
375	the craft	21.79	133	NP	physical entities (vessel)
376	and the crew	24.39	133	other NP fragment	agent
377	unlikely that	718.92	132	THAT CLAUSE	stance
378	emergency drills	1156.37	132	NP	notion

379	the officer	224.54	132	NP	agent
380	carried on board	937.19	132	VP-P	acvity/action
381	falling overboard	1492.55	132	VP-A	acvity/action
382	required that	163.01	131	THAT CLAUSE	stance
383	bilge pump	1270.27	131	NP	physical entities (equipment)
384	the dredge	182.64	130	NP	physical entities (vessel)
385	the gear	11.53	130	NP	physical entities (equipment)
386	the bow thruster	179.29	130	NP	physical entities (equipment)
387	to the deck	144.25	129	PP-based fragment	place reference
388	the buoy	58.69	129	NP	physical entities (equipment)
389	lifting equipment	1072.65	128	NP	physical entities (equipment)
390	is therefore	520.94	128	BE+ADJ./NOUN/PP	discourse organizer
391	the wreck	240.43	127	NP	place reference
392	merchant shipping and fishing vessels	1513.46	127	NP	physical entities (vessel)
393	the machinery	56.12	127	NP	physical entities (equipment)
394	sea survival	1076.04	127	NP	notion
395	result in	345.69	127	VP-A	discourse organizer
396	fitted in	171.55	127	VP-P	acvity/action
397	to the wheelhouse	138.05	126	PP-based fragment	place reference
398	by the skipper	264.46	126	PP-based fragment	agent
399	the second engineer	603.08	126	NP	agent
400	list of	320.04	125	NP+of fragment	specification of attributes
401	realised that	742.47	125	THAT CLAUSE	discourse organizer
402	with the requirements of	581.87	124	PP-based fragment	specification of attributes
403	there was no evidence	1210.46	124	SENTENCE STEM	notion
404	used as	330.55	124	VP-P	discourse organizer
405	as amended	808.30	124	ADV fragment	discourse organizer
406	the lifting	45.71	124	NP	acvity/action
407	was clear	237.60	123	BE+ADJ./NOUN/PP	stance
408	the tanker	153.40	123	NP	physical entities (vessel)
409	switched on	612.05	123	VP-P	acvity/action
410	appeared to be	832.02	122	VP-A	stance
411	at a speed of	657.30	122	PP-based fragment	specification of attributes
412	fishing grounds	1130.70	122	NP	place reference
413	general cargo	913.08	122	NP	physical entities (equipment)
414	standing orders	1554.76	122	NP	notion
415	the propulsion	60.62	122	NP	acvity/action
416	requires that	529.11	121	THAT CLAUSE	stance
417	effectiveness of	471.06	121	NP+of fragment	stance
418	of the hull	299.02	121	PP-based fragment	place reference
419	the trawl	118.68	121	NP	physical entities (equipment)
420	examination of	278.61	121	NP+of fragment	acvity/action
421	awareness of	244.62	120	NP+of fragment	stance

422	bilge alarm	1083.64	120	NP	physical entities (equipment)
423	the load	23.98	120	NP	notion
424	full astern	1138.43	120	ADJ. fragment	notion
425	clipper point	1290.67	119	NP	physical entities (equipment)
426	the bulwark	147.67	119	NP	physical entities (equipment)
427	and chief officer	551.16	119	other NP fragment	agent
428	deck officers	781.65	119	NP	agent
429	to secure	412.99	119	TO CLAUSE	acvity/action
430	search and rescue	1526.28	119	NP	acvity/action
431	was not possible	570.47	118	BE+ADJ./NOUN/PP	stance
432	the workboat	155.49	118	NP	physical entities (vessel)
433	for the crew	196.12	118	PP-based fragment	agent
434	safe operation of	751.45	117	NP+of fragment	stance
435	main engines	1079.55	116	NP	physical entities (equipment)
436	consisted of	484.89	116	VP-A	discourse organizer
437	hatch cover	1324.15	115	NP	place reference
438	vessel would	125.42	114	SENTENCE STEM	stance
439	watertight doors	1362.90	114	NP	place reference
440	watch alarm	887.78	114	NP	physical entities (equipment)
441	maib investigation	949.10	114	NP	notion
442	identified as	350.98	114	VP-P	discourse organizer
443	advised that	492.53	114	THAT CLAUSE	discourse organizer
444	attempting to	453.74	113	VP-A	stance
445	of vessel	293.92	113	PP-based fragment	physical entities (vessel)
446	its fleet	720.50	113	NP	physical entities (vessel)
447	passenger vessels	598.83	113	NP	physical entities (vessel)
448	limited to	150.48	113	VP-P	discourse organizer
449	maintenance of	132.88	113	NP+of fragment	acvity/action
450	the position of	279.18	112	NP+of fragment	specification of attributes
451	held an stcw	1404.89	112	VP-A	regulation
452	the port and starboard	810.48	112	NP	place reference
453	the wheel	114.89	112	NP	physical entities (equipment)
454	his crew	254.10	112	NP	agent
455	avoiding action	1303.91	112	VP-A	acvity/action
456	it would be	553.89	111	IT CLAUSE	stance
457	is unlikely	712.18	111	BE+ADJ./NOUN/PP	stance
458	from the bridge	320.84	111	PP-based fragment	place reference
459	full ahead	993.33	111	ADJ. fragment	notion
460	the pilots	37.40	111	NP	agent
461	to berth	149.37	111	PP-based fragment	acvity/action
462	the carriage	149.68	111	NP	acvity/action
463	are required to	521.16	110	VP-P	stance
464	was responsible	397.15	110	BE+ADJ./NOUN/PP	stance

465	to vessels	38.29	110	PP-based fragment	physical entities (vessel)
466	paper charts	1461.43	110	NP	physical entities (equipment)
467	recommended that	346.65	110	THAT CLAUSE	discourse organizer
468	been fitted	359.17	110	VP-P	activity/action
469	forward of	74.11	109	NP+of fragment	place reference
470	the compartment	43.36	109	NP	place reference
471	the engines	56.10	109	NP	physical entities (equipment)
472	length overall	1242.82	109	NP	notion
473	was found to be	620.79	109	VP-P	discourse organizer
474	continue to	320.50	109	VP-A	discourse organizer
475	to manoeuvre	275.58	109	TO CLAUSE	activity/action
476	an attempt to	584.28	108	other NP fragment	stance
477	safe working practices	1272.77	108	NP	regulation
478	on the starboard side of	727.74	108	PP-based fragment	place reference
479	container ship	798.74	108	NP	physical entities (vessel)
480	the submarine	145.19	108	NP	physical entities (vessel)
481	maib inspectors	1222.72	108	NP	agent
482	secured to	190.73	108	VP-P	activity/action
483	marine guidance	698.43	107	NP	regulation
484	of the deck	97.28	107	PP-based fragment	place reference
485	the towline	180.97	107	NP	physical entities (equipment)
486	conjunction with	776.08	107	other NP fragment	discourse organizer
487	relevant to	203.89	107	ADJ. fragment	discourse organizer
488	the chief officer and	391.86	107	other NP fragment	agent
489	its crew	217.68	107	NP	agent
490	safe navigation	787.92	106	NP	stance
491	in the engine room	630.76	106	PP-based fragment	place reference
492	his cabin	705.04	106	NP	place reference
493	the fleet	25.49	106	NP	physical entities (vessel)
494	a fishing vessel	487.16	106	NP	physical entities (vessel)
495	in respect of	505.24	106	PP-based fragment	discourse organizer
496	third engineer	877.93	106	NP	agent
497	a crewman	297.67	106	NP	agent
498	the harbourmaster	153.38	106	NP	agent
499	the autopilot	153.38	106	NP	agent
500	would also	296.69	105	other VP-based fragment	stance
501	the emergency services	669.77	105	NP	physical entities (equipment)
502	collision avoidance	1216.20	105	NP	notion
503	in conjunction with	666.26	105	PP-based fragment	discourse organizer
504	to alert	246.95	105	TO CLAUSE	activity/action
505	the safe operation	593.44	104	NP	stance
506	were unable	507.60	104	BE+ADJ./NOUN/PP	stance
507	were unable to	470.00	104	BE+ADJ./NOUN/PP	stance



508	chain locker	1400.52	104	NP	place reference
509	the chart plotter	773.62	104	NP	physical entities (equipment)
510	considered that	275.78	104	THAT CLAUSE	discourse organizer
511	regardless of	428.08	104	PP-based fragment	discourse organizer
512	crew on board	575.17	104	NP	agent
513	sailed from	653.79	104	VP-A	acvity/action
514	the waterline	212.63	104	NP	physical entities (others)
515	it was not possible	778.92	103	IT CLAUSE	stance
516	the rudder	105.98	103	NP	physical entities (equipment)
517	to sail	294.04	103	TO CLAUSE	acvity/action
518	heading of	158.65	103	NP+of fragment	acvity/action
519	risks associated with	933.54	102	other NP fragment	notion
520	was therefore	156.46	102	BE+ADJ./NOUN/PP	discourse organizer
521	applicable to	299.08	102	ADJ. fragment	discourse organizer
522	the conduct of	296.59	102	NP+of fragment	acvity/action
523	the carriage of	300.10	102	NP+of fragment	acvity/action
524	was secured	278.68	102	VP-P	activity/action
525	control room	560.14	101	NP	place reference
526	the generator	63.36	101	NP	physical entities (equipment)
527	the pec holder	170.28	101	NP	physical entities (equipment)
528	of crew	51.88	101	PP-based fragment	agent
529	passengers and crew	905.82	101	NP	agent
530	in an attempt to	474.12	100	PP-based fragment	stance
531	clear that	254.42	99	THAT CLAUSE	stance
532	was responsible for	517.29	99	BE+ADJ./NOUN/PP	stance
533	the deck and	79.44	99	other NP fragment	place reference
534	marine services	848.77	99	NP	notion
535	to be fitted	336.11	99	TO CLAUSE	acvity/action
536	is likely to	400.34	98	BE+ADJ./NOUN/PP	stance
537	unlikely to	267.74	98	ADJ. fragment	stance
538	electronic chart	1098.08	98	NP	notion
539	the route	24.00	98	NP	notion
540	to release	164.93	98	TO CLAUSE	acvity/action
541	on board for	331.26	98	PP-based fragment	acvity/action
542	the effectiveness of	294.38	97	NP+of fragment	stance
543	capacity of	280.65	97	NP+of fragment	specification of attributes
544	to port and	172.97	97	PP-based fragment	place reference
545	upper deck	787.04	97	NP	place reference
546	cargo ship	388.23	97	NP	physical entities (vessel)
547	of the starboard	67.85	96	PP-based fragment	place reference
548	recreational craft	1133.57	96	NP	physical entities (vessel)
549	tow line	862.35	96	NP	physical entities (equipment)
550	bridge procedures	491.23	96	NP	notion

551	in turn	221.64	96	PP-based fragment	discourse organizer
552	the engineers	62.76	96	NP	agent
553	not require	486.01	95	VP-A	stance
554	the weight of	243.43	95	NP+of fragment	specification of attributes
555	on the port side of	626.83	95	PP-based fragment	place reference
556	both vessels	349.67	95	NP	physical entities (vessel)
557	confirm that	499.84	95	THAT CLAUSE	discourse organizer
558	with respect to	534.44	95	PP-based fragment	discourse organizer
559	is evident that	605.13	94	THAT CLAUSE	stance
560	was unaware	387.94	94	BE+ADJ./NOUN/PP	stance
561	the results of	344.34	94	NP+of fragment	specification of attributes
562	in the ship	62.44	94	PP-based fragment	physical entities (vessel)
563	abandon ship	863.87	94	NP	physical entities (vessel)
564	the ballast	28.69	94	NP	physical entities (equipment)
565	the chain locker	776.75	94	NP	physical entities (equipment)
566	assumed that	490.85	94	THAT CLAUSE	discourse organizer
567	and although	17.29	94	ADV fragment	discourse organizer
568	bridge teams	809.94	94	NP	agent
569	also required	314.06	93	VP-A	stance
570	dry dock	1187.80	93	NP	place reference
571	aft deck	501.32	93	NP	place reference
572	a cargo	33.82	93	NP	physical entities (equipment)
573	the wire	13.40	93	NP	physical entities (equipment)
574	the vhf radio	10.19	93	NP	physical entities (equipment)
575	working on deck	776.02	93	VP-A	notion
576	of shipping	91.93	93	PP-based fragment	notion
577	fishing vessel safety	678.21	93	NP	notion
578	was equipped with	527.10	93	VP-P	acvity/action
579	the launch	78.07	93	NP	acvity/action
580	imo resolution	1206.59	92	NP	regulation
581	from the wheelhouse	355.89	92	PP-based fragment	place reference
582	steering gear	863.59	92	NP	physical entities (equipment)
583	established that	389.77	92	THAT CLAUSE	discourse organizer
584	it is therefore	669.74	92	IT CLAUSE	discourse organizer
585	with the master	79.00	92	PP-based fragment	agent
586	the crew and	36.99	92	other NP fragment	agent
587	the cadets	111.26	92	NP	agent
588	officer of the watch	1049.60	92	NP	agent
589	coast of	265.83	91	NP+of fragment	place reference
590	propeller pitch	945.43	91	NP	physical entities (equipment)
591	registered length	893.65	91	NP	notion
592	was used to	254.86	91	VP-P	discourse organizer
593	fitting of	263.97	91	NP+of fragment	acvity/action

594	the release	17.48	91	NP	acvity/action
595	high water	468.47	91	NP	physical entities (others)
596	was not possible to	384.81	90	BE+ADJ./NOUN/PP	stance
597	was based on	380.59	90	VP-P	discourse organizer
598	by the maib	330.19	90	PP-based fragment	agent
599	the skippers	28.70	90	NP	agent
600	on the ship	85.38	89	PP-based fragment	physical entities (vessel)
601	a fleet	212.33	89	NP	physical entities (vessel)
602	the container ship	475.01	89	NP	physical entities (vessel)
603	the hook	58.77	89	NP	physical entities (equipment)
604	means of navigation	1052.82	89	NP	notion
605	was subsequently	312.47	89	BE+ADJ./NOUN/PP	discourse organizer
606	vessel owners	355.45	89	NP	agent
607	port state	531.66	89	NP	agent
608	collision with	261.86	89	other NP fragment	acvity/action
609	permit to	296.34	88	VP-P	stance
610	it was not possible to	513.14	88	IT CLAUSE	stance
611	it is evident that	584.68	88	IT CLAUSE	stance
612	is clear	310.24	88	BE+ADJ./NOUN/PP	stance
613	evident from	521.78	88	ADJ. fragment	stance
614	for fishing vessels	464.15	88	PP-based fragment	physical entities (vessel)
615	the ship and	68.96	88	other NP fragment	physical entities (vessel)
616	aft mooring	703.07	88	NP	physical entities (equipment)
617	engine control	355.45	88	NP	physical entities (equipment)
618	maritime and coastguard agency	1114.02	88	NP	agent
619	corrective action	1021.86	88	NP	acvity/action
620	the pilotage	18.63	88	NP	acvity/action
621	the vessel would	133.16	87	SENTENCE STEM	stance
622	is no evidence	524.13	87	BE+ADJ./NOUN/PP	stance
623	bottom of	237.84	87	NP+of fragment	specification of attributes
624	to the starboard	48.76	87	PP-based fragment	place reference
625	fleet of	124.13	87	NP+of fragment	physical entities (vessel)
626	on the chart	382.82	87	PP-based fragment	physical entities (equipment)
627	propulsion control	667.33	87	NP	physical entities (equipment)
628	proper lookout	1002.62	87	NP	agent
629	to rest	130.50	87	TO CLAUSE	acvity/action
630	course in	87.17	87	other NP fragment	acvity/action
631	a heading	191.64	87	NP	acvity/action
632	the fitting	85.05	87	NP	acvity/action
633	bridge watchkeeping	639.80	87	NP	acvity/action
634	good practice	743.14	86	NP	stance
635	was aware of	260.29	86	BE+ADJ./NOUN/PP	stance
636	would not have been	571.31	86	BE+ADJ./NOUN/PP	stance

637	by the ship	164.75	86	PP-based fragment	physical entities (vessel)
638	safety officer	196.43	86	NP	notion
639	probable that	510.83	86	THAT CLAUSE	discourse organizer
640	all crew	199.36	86	NP	agent
641	a proper lookout	763.03	86	NP	agent
642	to alter	314.69	86	TO CLAUSE	acvity/action
643	been on board	120.31	86	BE+ADJ./NOUN/PP	acvity/action
644	in the absence of	335.65	85	PP-based fragment	specification of attributes
645	fishing vessels and	345.00	85	other NP fragment	physical entities (vessel)
646	the hauler	146.66	85	NP	physical entities (equipment)
647	the creels	68.19	85	NP	physical entities (equipment)
648	the net drum	140.68	85	NP	physical entities (equipment)
649	was reported to	280.63	85	VP-P	discourse organizer
650	have resulted in	456.92	85	VP-A	discourse organizer
651	recognised that	313.20	85	THAT CLAUSE	discourse organizer
652	port authority	484.59	85	NP	agent
653	course to port	667.13	85	NP	acvity/action
654	more likely	592.78	84	ADJ. fragment	stance
655	at work regulations	589.65	84	NP	regulation
656	to starboard and	232.99	84	PP-based fragment	place reference
657	from the port	191.95	84	PP-based fragment	place reference
658	engine speed	382.33	84	NP	notion
659	to indicate that	347.01	84	THAT CLAUSE	discourse organizer
660	proceed to	244.93	84	VP-A	acvity/action
661	passed to	109.81	84	VP-A	acvity/action
662	on duty	241.79	84	PP-based fragment	acvity/action
663	bulk cargoes	1047.85	84	NP	physical entities (others)
664	low water	559.49	84	NP	physical entities (others)
665	be required	122.66	83	VP-P	stance
666	code of safe working	957.76	83	NP	regulation
667	passenger vessel	268.60	83	NP	physical entities (vessel)
668	two vessels	240.49	83	NP	physical entities (vessel)
669	on vhf radio	234.45	83	PP-based fragment	physical entities (equipment)
670	vhf radio and	550.38	83	other NP fragment	physical entities (equipment)
671	typhoon clipper	1144.21	83	NP	physical entities (equipment)
672	the centreline	146.97	83	NP	physical entities (equipment)
673	thermal oil	999.84	83	NP	physical entities (equipment)
674	the risks associated	539.83	83	other NP fragment	notion
675	ship management	360.51	83	NP	notion
676	chief engineer and	426.56	83	other NP fragment	agent
677	had sailed	423.82	83	VP-A	acvity/action
678	fresh water	746.12	83	NP	physical entities (others)
679	wave height	933.14	83	NP	physical entities (others)

680	while the vessel was	474.31	82	ADV fragment	time reference
681	not required to	275.25	82	VP-P	stance
682	he was unable to	433.03	82	SENTENCE STEM	stance
683	it is unlikely	600.37	82	IT CLAUSE	stance
684	likely to be	444.98	82	ADJ. fragment	stance
685	imsbc code	908.67	82	NP	regulation
686	operations manual	591.55	82	NP	regulation
687	port quarter	668.41	82	NP	place reference
688	engine compartment	574.55	82	NP	place reference
689	the upper deck	420.23	82	NP	place reference
690	of the vessels	161.74	82	PP-based fragment	physical entities (vessel)
691	personal flotation	1103.15	82	NP	physical entities (equipment)
692	the risks associated with	574.90	82	other NP fragment	notion
693	the third engineer	149.51	82	NP	agent
694	master and chief officer	686.80	82	NP	agent
695	to lift	141.30	82	TO CLAUSE	activity/action
696	the keel	80.65	82	NP	physical entities (others)
697	the bottom of	225.62	81	NP+of fragment	specification of attributes
698	machinery space	786.71	81	NP	place reference
699	of ships	33.56	81	PP-based fragment	physical entities (vessel)
700	celtic carrier	1137.56	81	NP	physical entities (vessel)
701	the foam	22.39	81	NP	physical entities (equipment)
702	the valve	21.17	81	NP	physical entities (equipment)
703	there is no evidence	788.90	81	SENTENCE STEM	notion
704	vessel traffic	365.52	81	NP	notion
705	a mayday	337.42	81	NP	notion
706	the officer of	202.76	81	NP+of fragment	agent
707	other crew	210.09	81	NP	agent
708	classification societies	1176.26	81	NP	agent
709	the installation	59.58	81	NP	activity/action
710	planned maintenance	767.02	81	NP	activity/action
711	on completion of	340.56	80	PP-based fragment	time reference
712	should include	519.53	80	VP-A	stance
713	was apparent	218.76	80	BE+ADJ./NOUN/PP	stance
714	load of	84.75	80	NP+of fragment	specification of attributes
715	mess room	849.47	80	NP	place reference
716	the cargo hold	408.80	80	NP	place reference
717	the boat and	80.48	80	other NP fragment	physical entities (vessel)
718	buoyancy foam	895.91	80	NP	physical entities (equipment)
719	harbour authorities	782.65	80	NP	agent
720	alteration of	231.34	80	NP+of fragment	activity/action
721	the length of	206.57	79	NP+of fragment	specification of attributes
722	to the engine room	440.54	79	PP-based fragment	place reference

723	on board the vessel	278.02	79	PP-based fragment	physical entities (vessel)
724	of the vessel and	106.84	79	PP-based fragment	physical entities (vessel)
725	the trolley	127.57	79	NP	physical entities (equipment)
726	the lifejacket	19.55	79	NP	physical entities (equipment)
727	the deckhands	64.10	79	NP	agent
728	to mitigate	303.76	79	TO CLAUSE	acvity/action
729	the capsize	54.31	79	NP	acvity/action
730	timor stream	1132.50	79	NP	physical entities (others)
731	intended for	212.35	78	VP-P	stance
732	the fairway	139.26	78	NP	place reference
733	the hatch covers	562.09	78	NP	place reference
734	on fishing vessels	393.99	78	PP-based fragment	physical entities (vessel)
735	frequency vhf radio	930.48	78	NP	physical entities (equipment)
736	resulting from	417.28	78	VP-A	discourse organizer
737	decided that	306.32	78	THAT CLAUSE	discourse organizer
738	altered course	796.29	78	VP-A	acvity/action
739	berth at	270.13	78	VP-A	acvity/action
740	of the grounding	161.77	78	PP-based fragment	acvity/action
741	pitch control	552.99	78	NP	acvity/action
742	deck cargo	291.93	78	NP	physical entities (others)
743	warping drum	1056.77	77	NP	physical entities (equipment)
744	accommodation ladder	838.07	77	NP	physical entities (equipment)
745	radio channel	663.04	77	NP	physical entities (equipment)
746	fire pump	480.35	77	NP	physical entities (equipment)
747	the cylinder	49.27	77	NP	physical entities (equipment)
748	hazards associated	795.79	77	other NP fragment	notion
749	were found to	304.46	77	VP-P	discourse organizer
750	noticed that	389.41	77	THAT CLAUSE	discourse organizer
751	the harbour authority	437.69	77	NP	agent
752	not fitted with	411.27	77	VP-P	acvity/action
753	left the bridge	521.09	77	VP-A	acvity/action
754	a heading of	343.06	77	NP+of fragment	acvity/action
755	the fitting of	229.93	77	NP+of fragment	acvity/action
756	the securing	8.96	77	NP	acvity/action
757	is required to	250.37	76	VP-P	stance
758	hazards associated with	698.27	76	other NP fragment	stance
759	appropriate to	33.51	76	ADJ. fragment	stance
760	on the seabed	341.62	76	PP-based fragment	place reference
761	the aft deck	143.66	76	NP	place reference
762	port marine safety	635.04	76	NP	notion
763	to confirm that	344.65	76	THAT CLAUSE	discourse organizer
764	in response to	300.26	76	PP-based fragment	discourse organizer
765	the flag state	534.16	76	NP	agent



766	alter course	821.42	76	VP-A	acvity/action
767	to pull	186.42	76	TO CLAUSE	acvity/action
768	passage from	220.14	76	other NP fragment	acvity/action
769	was evident	238.24	75	BE+ADJ./NOUN/PP	stance
770	would probably	454.86	75	ADV fragment	stance
771	port bow	361.53	75	NP	place reference
772	the rescue boat	379.23	75	NP	physical entities (vessel)
773	the cable	33.81	75	NP	physical entities (equipment)
774	control lever	646.97	75	NP	physical entities (equipment)
775	the fishing gear	369.34	75	NP	physical entities (equipment)
776	subjected to	307.41	75	VP-P	discourse organizer
777	international maritime organization	1065.41	75	NP	agent
778	listed in	283.74	75	VP-P	acvity/action
779	is intended	259.33	74	VP-P	stance
780	was intended to	244.41	74	VP-P	stance
781	would have had	95.28	74	other VP-based fragment	stance
782	would also have	577.88	74	other VP-based fragment	stance
783	intention to	219.22	74	other NP fragment	stance
784	no evidence of	374.18	74	NP+of fragment	stance
785	it was apparent	565.90	74	IT CLAUSE	stance
786	is apparent	360.95	74	BE+ADJ./NOUN/PP	stance
787	likely to have been	515.90	74	ADJ. fragment	stance
788	the cockpit	124.75	74	NP	place reference
789	had been on board	508.34	74	BE+ADJ./NOUN/PP	place reference
790	audible alarm	803.02	74	NP	physical entities (equipment)
791	vhf radio channel	808.05	74	NP	physical entities (equipment)
792	therefore it is	530.18	74	IT CLAUSE	discourse organizer
793	paper chart	808.71	74	NP	physical entities (others)
794	was aware that	328.65	73	THAT CLAUSE	stance
795	it is apparent	564.80	73	IT CLAUSE	stance
796	effect on	190.04	73	other NP fragment	specification of attributes
797	either side of	458.10	73	NP+of fragment	place reference
798	ballast tank	702.91	73	NP	place reference
799	mooring lines	729.71	73	NP	physical entities (equipment)
800	skippers of	113.26	73	NP+of fragment	agent
801	service engineers	691.92	73	NP	agent
802	the watchkeeper	64.79	73	NP	agent
803	fitted with an	330.74	73	VP-P	acvity/action
804	testing of	136.33	73	NP+of fragment	acvity/action
805	the discharge	18.51	73	NP	acvity/action
806	before the collision	554.52	72	PP-based fragment	time reference
807	no evidence that	460.51	72	THAT CLAUSE	stance
808	was safe	40.51	72	BE+ADJ./NOUN/PP	stance

809	the forepeak	139.24	72	NP	place reference
810	bow thrusters	848.25	72	NP	physical entities (equipment)
811	the engine control	242.46	72	NP	physical entities (equipment)
812	protective equipment	713.38	72	NP	physical entities (equipment)
813	steering system	452.46	72	NP	notion
814	crew training	204.55	72	NP	notion
815	marine safety code	769.97	72	NP	notion
816	prepared to	153.58	72	VP-P	discourse organizer
817	referred to as	413.82	72	VP-P	discourse organizer
818	check that	230.12	72	THAT CLAUSE	discourse organizer
819	service engineer	452.27	72	NP	agent
820	mca surveyors	594.41	72	NP	agent
821	held on board	491.34	72	VP-P	acvity/action
822	to the collision	92.52	72	PP-based fragment	acvity/action
823	ship handling	503.13	72	NP	acvity/action
824	the manoeuvre	26.28	72	NP	acvity/action
825	the manoeuvring	11.54	72	NP	acvity/action
826	essential that	313.59	71	THAT CLAUSE	stance
827	is essential	401.27	71	BE+ADJ./NOUN/PP	stance
828	not aware of	286.23	71	ADJ. fragment	stance
829	solas chapter	857.79	71	NP	regulation
830	deck of	19.72	71	NP+of fragment	place reference
831	astern pitch	668.32	71	NP	physical entities (equipment)
832	awareness course	581.20	71	NP	notion
833	the buoyancy	10.64	71	NP	notion
834	slow ahead	745.55	71	ADJ. fragment	notion
835	it is considered	494.14	71	IT CLAUSE	discourse organizer
836	and the pilot	28.88	71	other NP fragment	agent
837	and the chief officer	315.22	71	other NP fragment	agent
838	monitoring of	102.02	71	NP+of fragment	acvity/action
839	marine operations	421.34	71	NP	acvity/action
840	the list	16.50	71	NP	acvity/action
841	alteration of course	928.69	71	NP	acvity/action
842	speed of knots	747.08	71	NP	physical entities (others)
843	crew would	104.93	70	SENTENCE STEM	stance
844	it is probable	542.92	70	IT CLAUSE	stance
845	is probable	481.67	70	BE+ADJ./NOUN/PP	stance
846	the height of	169.83	70	NP+of fragment	specification of attributes
847	small vessels	307.38	70	NP	physical entities (vessel)
848	shell plating	975.08	70	NP	physical entities (equipment)
849	a crew member	589.71	70	NP	agent
850	worked on board	510.37	70	VP-A	acvity/action
851	is probable that	478.07	69	THAT CLAUSE	stance

852	the imsb code	449.56	69	NP	regulation
853	port bridge	165.12	69	NP	place reference
854	the offshore	12.62	69	NP	place reference
855	starboard bow	423.66	69	NP	place reference
856	fishing vessels of	283.83	69	NP+of fragment	physical entities (vessel)
857	commercial vessel	279.17	69	NP	physical entities (vessel)
858	auxiliary engine	602.28	69	NP	physical entities (equipment)
859	propulsion system	457.29	69	NP	physical entities (equipment)
860	personal protective equipment	859.31	69	NP	physical entities (equipment)
861	have caused	321.53	69	VP-A	discourse organizer
862	port state control	669.55	69	NP	agent
863	her crew	238.53	69	NP	agent
864	to navigate	264.73	69	TO CLAUSE	acvity/action
865	the wearing of	200.76	69	NP+of fragment	acvity/action
866	the berthing	51.32	69	NP	acvity/action
867	electrical power	647.12	69	NP	physical entities (others)
868	when the master	120.35	68	ADV fragment	time reference
869	contributory factor	868.39	68	NP	stance
870	high risk	342.23	68	NP	stance
871	stern of	41.82	68	NP+of fragment	place reference
872	each side	388.32	68	NP	place reference
873	of the harbour	105.01	68	PP-based fragment	place reference
874	of the hold	179.25	68	PP-based fragment	place reference
875	from the engine	155.99	68	PP-based fragment	physical entities (equipment)
876	kill cord	1111.43	68	NP	physical entities (equipment)
877	immersion suits	1007.12	68	NP	physical entities (equipment)
878	the steel	15.48	68	NP	physical entities (equipment)
879	the aft mooring	123.90	68	NP	physical entities (equipment)
880	steering control	462.43	68	NP	physical entities (equipment)
881	means of access	870.53	68	NP	notion
882	vessels safety	86.01	68	NP	notion
883	acting on	325.99	68	VP-A	discourse organizer
884	stowed in	229.57	68	VP-P	acvity/action
885	release of	81.61	68	NP+of fragment	acvity/action
886	was not aware	312.62	67	BE+ADJ./NOUN/PP	stance
887	was no requirement	297.65	67	BE+ADJ./NOUN/PP	stance
888	loading of	39.44	67	NP+of fragment	specification of attributes
889	from the deck	220.29	67	PP-based fragment	place reference
890	at the port	90.16	67	PP-based fragment	place reference
891	the stern of	122.48	67	NP+of fragment	place reference
892	vessel of	431.58	67	NP+of fragment	physical entities (vessel)
893	bulk carrier	821.79	67	NP	physical entities (vessel)
894	ballast tanks	729.99	67	NP	physical entities (vessel)

895	a liferaft	108.42	67	NP	physical entities (vessel)
896	the brake	55.19	67	NP	physical entities (equipment)
897	course and speed	706.33	67	NP	notion
898	agreed that	263.01	67	THAT CLAUSE	discourse organizer
899	in this report	288.52	67	PP-based fragment	discourse organizer
900	the cook	82.04	67	NP	agent
901	fitted on board	407.29	67	VP-P	acvity/action
902	float free	844.48	67	VP-A	acvity/action
903	was heading	123.02	67	VP-A	acvity/action
904	transfer of	109.22	67	NP+of fragment	acvity/action
905	lifting operations	520.49	67	NP	acvity/action
906	hydraulic oil	649.08	67	NP	physical entities (others)
907	the potential to	118.75	66	other NP fragment	stance
908	the stcw	8.10	66	NP	regulation
909	forward end	511.09	66	NP	place reference
910	speed craft	453.92	66	NP	physical entities (vessel)
911	high speed craft	756.17	66	NP	physical entities (vessel)
912	the gangway	98.97	66	NP	physical entities (equipment)
913	safety of small fishing	645.40	66	NP	notion
914	reported to be	353.21	66	VP-P	discourse organizer
915	coastguard agency is recommended	849.49	66	SENTENCE STEM	agent
916	the master and chief officer	454.26	66	NP	agent
917	the motorman	108.28	66	NP	agent
918	wearing a lifejacket	801.28	66	VP-A	acvity/action
919	arrived on the bridge	582.71	66	VP-A	acvity/action
920	emergency release	502.30	66	NP	acvity/action
921	a load	107.48	66	NP	acvity/action
922	a rescue	75.28	66	NP	acvity/action
923	should ensure	289.25	65	VP-A	stance
924	fortunate that	396.65	65	THAT CLAUSE	stance
925	it is probable that	456.16	65	IT CLAUSE	stance
926	the primary means of	184.14	65	NP+of fragment	specification of attributes
927	an angle	387.40	65	NP	specification of attributes
928	deck level	347.44	65	NP	place reference
929	the port bridge	185.50	65	NP	place reference
930	the quayside	129.63	65	NP	place reference
931	the fishing grounds	414.33	65	NP	place reference
932	cargo vessel	82.18	65	NP	physical entities (vessel)
933	of the crane	128.85	65	PP-based fragment	physical entities (equipment)
934	conveyor belt	946.97	65	NP	physical entities (equipment)
935	flotation devices	929.37	65	NP	physical entities (equipment)
936	the gearbox	53.57	65	NP	physical entities (equipment)
937	the main engines	340.53	65	NP	physical entities (equipment)

938	bridge resource	570.09	65	NP	notion
939	the weather conditions	316.87	65	NP	notion
940	work and rest	715.03	65	NP	notion
941	see section	584.56	65	VP-A	discourse organizer
942	indicating that	332.44	65	THAT CLAUSE	discourse organizer
943	the implementation of	176.84	65	NP+of fragment	acvity/action
944	cargo securing	462.15	65	NP	acvity/action
945	strength of	138.91	64	NP+of fragment	specification of attributes
946	the direction of	175.11	64	NP+of fragment	specification of attributes
947	of the hatch	135.89	64	NP+of fragment	place reference
948	the coast	41.44	64	NP	place reference
949	the engine compartment	307.49	64	NP	place reference
950	towing hook	683.97	64	NP	physical entities (equipment)
951	salvage pump	672.00	64	NP	physical entities (equipment)
952	the gantry	70.56	64	NP	physical entities (equipment)
953	the deficiencies	10.13	64	NP	notion
954	this resulted in	326.07	64	SENTENCE STEM	discourse organizer
955	the captain	50.88	64	NP	agent
956	technical superintendent	741.90	64	NP	agent
957	collided with	407.35	64	VP-A	acvity/action
958	joined the vessel	360.42	64	VP-A	acvity/action
959	to steer	254.38	64	TO CLAUSE	acvity/action
960	removal of	232.21	64	NP+of fragment	acvity/action
961	installation of	137.38	64	NP+of fragment	acvity/action
962	is unlikely that	385.90	63	THAT CLAUSE	stance
963	reasonably practicable	963.24	63	ADJ. fragment	stance
964	highly likely	683.99	63	ADJ. fragment	stance
965	enclosed space	688.11	63	NP	place reference
966	cargo ships	341.14	63	NP	physical entities (vessel)
967	another vessel	258.64	63	NP	physical entities (vessel)
968	their vessel	53.27	63	NP	physical entities (vessel)
969	the engine and	31.00	63	other NP fragment	physical entities (equipment)
970	bilge pumps	676.73	63	NP	physical entities (equipment)
971	port engine	143.94	63	NP	physical entities (equipment)
972	an engine	98.19	63	NP	physical entities (equipment)
973	an epirb	362.84	63	NP	physical entities (equipment)
974	the angle	22.50	63	NP	notion
975	it was reported	385.85	63	IT CLAUSE	discourse organizer
976	to crew	123.39	63	PP-based fragment	agent
977	master and pilot	462.68	63	NP	agent
978	lead pilot	527.08	63	NP	agent
979	two deckhands	522.66	63	NP	agent
980	be maintained	283.25	63	VP-P	acvity/action

981	at anchor	204.29	63	PP-based fragment	acvity/action
982	back on board	288.45	63	PP-based fragment	acvity/action
983	securing of	73.06	63	NP+of fragment	acvity/action
984	practice of	22.53	63	NP+of fragment	acvity/action
985	cargo handling	475.76	63	NP	acvity/action
986	intended to be	302.46	62	VP-A	stance
987	it is unlikely that	427.90	62	IT CLAUSE	stance
988	is evident from	424.09	62	BE+ADJ./NOUN/PP	stance
989	of safe working practices	638.66	62	PP-based fragment	regulation
990	bridge procedures guide	712.06	62	NP	regulation
991	at the stern	256.57	62	PP-based fragment	place reference
992	of fuel	52.77	62	PP-based fragment	physical entities (others)
993	buoyancy aids	760.82	62	NP	physical entities (equipment)
994	anchor cable	645.86	62	NP	physical entities (equipment)
995	bilge alarms	602.89	62	NP	physical entities (equipment)
996	forward mooring	455.52	62	NP	physical entities (equipment)
997	the radar display	390.58	62	NP	physical entities (equipment)
998	in accordance with the requirements	595.36	62	PP-based fragment	notion
999	as follows	448.90	62	PP-based fragment	discourse organizer
1000	in line with	378.31	62	PP-based fragment	discourse organizer
1001	deck figure	159.69	62	NP	agent
1002	senior engineer	516.38	62	NP	agent
1003	deck crew	114.43	62	NP	agent
1004	was switched	198.99	62	VP-P	acvity/action
1005	be fitted with	286.98	62	VP-P	acvity/action
1006	be taken to	228.70	62	VP-P	acvity/action
1007	to anchor	43.10	62	TO CLAUSE	acvity/action
1008	solid bulk	816.24	62	NP	physical entities (others)
1009	radar target	634.36	62	NP	physical entities (others)
1010	should be taken	465.97	61	VP-P	stance
1011	been possible	181.31	61	BE+ADJ./NOUN/PP	stance
1012	was likely	62.36	61	BE+ADJ./NOUN/PP	stance
1013	of bridge	29.26	61	PP-based fragment	place reference
1014	void space	734.00	61	NP	place reference
1015	port bridge wing	691.97	61	NP	place reference
1016	container vessel	249.07	61	NP	physical entities (vessel)
1017	commercial vessels	326.36	61	NP	physical entities (vessel)
1018	mooring ropes	626.08	61	NP	physical entities (equipment)
1019	control panel	527.43	61	NP	physical entities (equipment)
1020	load line	497.94	61	NP	physical entities (equipment)
1021	general alarm	466.37	61	NP	physical entities (equipment)
1022	fire alarm	322.31	61	NP	physical entities (equipment)
1023	the joystick	76.67	61	NP	physical entities (equipment)



1024	fishing vessels safety	402.26	61	NP	notion
1025	resulted from	255.66	61	VP-A	discourse organizer
1026	watchkeeping officers	577.90	61	NP	agent
1027	deck officer	157.21	61	NP	agent
1028	was not fitted	286.92	61	VP-P	acvity/action
1029	was lifted	210.78	61	VP-P	acvity/action
1030	made contact with	473.55	61	VP-A	acvity/action
1031	master ordered	363.62	61	SENTENCE STEM	acvity/action
1032	consultation with	393.92	61	other NP fragment	acvity/action
1033	intended track	571.75	61	NP	acvity/action
1034	water level	315.28	61	NP	physical entities (others)
1035	to be completed	223.44	60	TO CLAUSE	stance
1036	it is clear	432.22	60	IT CLAUSE	stance
1037	were aware	197.78	60	BE+ADJ./NOUN/PP	stance
1038	in compliance with	306.47	60	PP-based fragment	specification of attributes
1039	vessel certificate	173.16	60	NP	regulation
1040	in the forward	118.20	60	PP-based fragment	place reference
1041	boiler room	653.36	60	NP	place reference
1042	inland waters	746.43	60	NP	place reference
1043	messenger line	716.32	60	NP	physical entities (vessel)
1044	propeller shaft	627.33	60	NP	physical entities (equipment)
1045	alarm system	279.91	60	NP	physical entities (equipment)
1046	if required	234.75	60	ADV fragment	discourse organizer
1047	for the master	9.41	60	PP-based fragment	agent
1048	the master or	80.49	60	other NP fragment	agent
1049	mca surveyor	463.64	60	NP	agent
1050	its masters	294.28	60	NP	agent
1051	the officer of the watch	627.45	60	NP	agent
1052	the lead pilot	124.51	60	NP	agent
1053	the port authority	247.07	60	NP	agent
1054	navigation in	60.95	60	other NP fragment	acvity/action
1055	was released	164.18	60	BE+ADJ./NOUN/PP	acvity/action
1056	was connected	150.70	60	BE+ADJ./NOUN/PP	acvity/action
1057	height of tide	764.66	60	NP	physical entities (others)
1058	would probably have	470.49	59	other VP-based fragment	stance
1059	appropriate for	119.01	59	ADJ. fragment	stance
1060	on each side	343.14	59	PP-based fragment	place reference
1061	mooring deck	308.72	59	NP	place reference
1062	other vessel	53.13	59	NP	physical entities (vessel)
1063	commodore clipper	825.84	59	NP	physical entities (equipment)
1064	the linkspan	98.88	59	NP	physical entities (equipment)
1065	visibility was good	711.74	59	SENTENCE STEM	notion
1066	assessed that	199.16	59	THAT CLAUSE	discourse organizer

1067	masters of	40.08	59	NP+of fragment	agent
1068	skipper and crew	369.90	59	NP	agent
1069	port captain	435.51	59	NP	agent
1070	manned by	337.76	59	VP-P	acvity/action
1071	to proceed to	193.47	59	TO CLAUSE	acvity/action
1072	master decided	334.90	59	SENTENCE STEM	acvity/action
1073	this included	155.07	58	SENTENCE STEM	stance
1074	adversely affected	858.60	58	ADJ. fragment	stance
1075	draught of	131.07	58	NP+of fragment	specification of attributes
1076	of solas	75.09	58	PP-based fragment	regulation
1077	responsibility of	69.45	58	NP+of fragment	regulation
1078	equipment regulations	326.32	58	NP	regulation
1079	at the forward	201.78	58	PP-based fragment	place reference
1080	watertight door	582.95	58	NP	place reference
1081	starboard quarter	526.99	58	NP	place reference
1082	of ship	43.79	58	PP-based fragment	physical entities (vessel)
1083	data recorder	740.03	58	NP	physical entities (equipment)
1084	dredge gear	599.06	58	NP	physical entities (equipment)
1085	mooring rope	450.28	58	NP	physical entities (equipment)
1086	starboard engine	203.32	58	NP	physical entities (equipment)
1087	the beam	10.48	58	NP	physical entities (equipment)
1088	chances of survival	934.79	58	NP	physical entities (equipment)
1089	voyage data recorder	888.34	58	NP	physical entities (equipment)
1090	personal flotation devices	838.03	58	NP	physical entities (equipment)
1091	the port engine	160.17	58	NP	physical entities (equipment)
1092	the tow line	88.11	58	NP	physical entities (equipment)
1093	meet the requirements	570.26	58	VP-A	notion
1094	of gravity	219.28	58	PP-based fragment	notion
1095	sail training	483.02	58	NP	notion
1096	bridge resource management	579.23	58	NP	notion
1097	respond to	179.52	58	VP-A	discourse organizer
1098	stating that	333.43	58	THAT CLAUSE	discourse organizer
1099	senior officers	559.03	58	NP	agent
1100	the engineer deckhand	376.16	58	NP	agent
1101	turn to starboard	546.67	58	VP-A	acvity/action
1102	board its vessels	465.15	58	VP-A	acvity/action
1103	of lifting	50.52	58	PP-based fragment	acvity/action
1104	course alteration	555.20	58	NP	acvity/action
1105	passed through	431.86	58	NP	acvity/action
1106	during the passage	372.10	57	PP-based fragment	time reference
1107	should be used	383.98	57	VP-P	stance
1108	to be conducted	194.31	57	TO CLAUSE	stance
1109	unaware that	251.31	57	THAT CLAUSE	stance

1110	requested that	224.64	57	THAT CLAUSE	stance
1111	it is essential	424.44	57	IT CLAUSE	stance
1112	the most likely	61.67	57	ADJ. fragment	stance
1113	working practices for	459.44	57	other NP fragment	regulation
1114	the forward end	313.70	57	NP	place reference
1115	of the engine room	277.30	57	PP-based fragment	place reference
1116	out of the water	450.78	57	NP	place reference
1117	of his vessel	104.99	57	PP-based fragment	physical entities (vessel)
1118	the tugs	10.67	57	NP	physical entities (vessel)
1119	the windlass	70.93	57	NP	physical entities (equipment)
1120	fire extinguishing system	553.75	57	NP	physical entities (equipment)
1121	the warping drum	444.95	57	NP	physical entities (equipment)
1122	in good condition	299.97	57	PP-based fragment	notion
1123	marine accident	221.56	57	NP	notion
1124	not considered to	233.27	57	VP-P	discourse organizer
1125	ensured that	331.83	57	THAT CLAUSE	discourse organizer
1126	compliant with	362.05	57	ADJ. fragment	discourse organizer
1127	royal yachting association	839.82	57	NP	agent
1128	owners and skippers	719.73	57	NP	agent
1129	a fisherman	222.38	57	NP	agent
1130	the senior engineer	333.41	57	NP	agent
1131	released from	290.14	57	VP-P	acvity/action
1132	had been fitted	348.03	57	VP-P	acvity/action
1133	proceeded to	177.79	57	VP-A	acvity/action
1134	for collision	100.53	57	PP-based fragment	acvity/action
1135	of the passage	78.65	57	PP-based fragment	acvity/action
1136	navigation of	12.43	57	NP+of fragment	acvity/action
1137	the course of	92.39	57	NP+of fragment	acvity/action
1138	skipper would	100.83	56	SENTENCE STEM	stance
1139	were unaware	292.99	56	BE+ADJ./NOUN/PP	stance
1140	is apparent that	336.34	56	BE+ADJ./NOUN/PP	stance
1141	was unaware of	205.29	56	BE+ADJ./NOUN/PP	stance
1142	to the forward	46.77	56	PP-based fragment	place reference
1143	freeing ports	758.53	56	NP	place reference
1144	enclosed spaces	667.72	56	NP	place reference
1145	tank top	498.98	56	NP	place reference
1146	leisure craft	635.45	56	NP	physical entities (vessel)
1147	registered fishing vessels	452.48	56	NP	physical entities (vessel)
1148	shaft generator	619.05	56	NP	physical entities (equipment)
1149	the ventilation	20.83	56	NP	physical entities (equipment)
1150	available on board	353.27	56	ADJ. fragment	notion
1151	stability assessment	342.91	56	NP	notion
1152	watertight integrity	717.17	56	NP	notion

1153	have prompted	391.17	56	VP-A	discourse organizer
1154	owing to	232.47	56	PP-based fragment	discourse organizer
1155	however although	222.36	56	ADV fragment	discourse organizer
1156	persons on board	447.60	56	NP	agent
1157	altering course	617.76	56	VP-A	acvity/action
1158	been wearing	230.15	56	VP-A	acvity/action
1159	to alter course	409.81	56	TO CLAUSE	acvity/action
1160	was on passage	209.90	56	BE+ADJ./NOUN/PP	acvity/action
1161	rough seas	728.71	56	NP	physical entities (others)
1162	the sea conditions	31.16	56	NP	physical entities (others)
1163	highly likely that	511.11	55	THAT CLAUSE	stance
1164	code requires	414.25	55	SENTENCE STEM	stance
1165	of safe working practices for	585.97	55	PP-based fragment	regulation
1166	fore and aft	652.50	55	NP	place reference
1167	the foredeck	108.38	55	NP	place reference
1168	open deck	291.21	55	NP	place reference
1169	small commercial vessel	497.35	55	NP	physical entities (vessel)
1170	to the surface	140.63	55	PP-based fragment	physical entities (others)
1171	cargo of	10.37	55	NP+of fragment	physical entities (others)
1172	power supply	503.45	55	NP	physical entities (equipment)
1173	for navigation	105.67	55	PP-based fragment	notion
1174	hazards of	54.02	55	NP+of fragment	notion
1175	the port marine safety	362.74	55	NP	notion
1176	for fishermen	153.94	55	PP-based fragment	agent
1177	a deckhand	95.98	55	NP	agent
1178	flag states	581.29	55	NP	agent
1179	lifeboat crew	255.06	55	NP	agent
1180	bridge watch	241.98	55	NP	agent
1181	be worn	308.67	55	VP-P	acvity/action
1182	was manned	202.37	55	VP-P	acvity/action
1183	to implement	177.41	55	TO CLAUSE	acvity/action
1184	to load	30.86	55	TO CLAUSE	acvity/action
1185	for collision avoidance	599.27	55	PP-based fragment	acvity/action
1186	courses in	140.84	55	other NP fragment	acvity/action
1187	alteration to	124.87	55	other NP fragment	acvity/action
1188	dry docking	749.52	55	NP	acvity/action
1189	life saving	737.55	55	NP	acvity/action
1190	towage operations	474.85	55	NP	acvity/action
1191	a passage	22.24	55	NP	acvity/action
1192	engine telegraph	486.37	55	NP	physical entities (others)
1193	was apparent that	255.04	54	THAT CLAUSE	stance
1194	unlikely to have	401.35	54	ADJ. fragment	stance
1195	necessary for	152.03	54	ADJ. fragment	stance

1196	possible for	94.28	54	ADJ. fragment	stance
1197	inspections of	65.06	54	NP+of fragment	specification of attributes
1198	the code of practice	419.92	54	NP	regulation
1199	alp forward	602.85	54	NP	place reference
1200	in the cabin	196.70	54	PP-based fragment	place reference
1201	the deck of	53.87	54	NP+of fragment	place reference
1202	the proximity of	149.26	54	NP+of fragment	place reference
1203	sea room	221.69	54	NP	place reference
1204	working deck	210.37	54	NP	place reference
1205	rescue boats	446.70	54	NP	physical entities (vessel)
1206	the two vessels	86.97	54	NP	physical entities (vessel)
1207	for cargo	33.10	54	PP-based fragment	physical entities (others)
1208	fall wire	540.27	54	NP	physical entities (equipment)
1209	the hopper	80.11	54	NP	physical entities (equipment)
1210	very high frequency (vhf) radio	659.58	54	NP	physical entities (equipment)
1211	bollard pull	781.48	54	NP	notion
1212	centre of gravity	737.39	54	NP	notion
1213	worked as	195.04	54	VP-P	discourse organizer
1214	were found to be	331.14	54	VP-P	discourse organizer
1215	of the crew were	146.07	54	PP-based fragment	agent
1216	avoid collision	472.41	54	VP-A	acvity/action
1217	to push	170.47	54	TO CLAUSE	acvity/action
1218	to pump	13.39	54	TO CLAUSE	acvity/action
1219	course to starboard	432.37	54	NP	acvity/action
1220	solid bulk cargoes	798.48	54	NP	physical entities (others)
1221	should be provided	438.12	53	VP-P	stance
1222	made aware	351.62	53	VP-A	stance
1223	there was no requirement for	431.83	53	SENTENCE STEM	stance
1224	a contributory factor	530.39	53	NP	stance
1225	it was apparent that	363.61	53	IT CLAUSE	stance
1226	is likely to have	360.79	53	BE+ADJ./NOUN/PP	stance
1227	more effective	359.83	53	ADJ. fragment	stance
1228	distance between	399.49	53	other NP fragment	specification of attributes
1229	the assistance of	161.17	53	NP+of fragment	specification of attributes
1230	on the forward	119.04	53	PP-based fragment	place reference
1231	on deck and	107.12	53	PP-based fragment	place reference
1232	in close proximity	516.68	53	PP-based fragment	place reference
1233	onto the deck	403.14	53	PP-based fragment	place reference
1234	escape hatch	590.31	53	NP	place reference
1235	harbour entrance	509.80	53	NP	place reference
1236	aft compartment	432.52	53	NP	place reference
1237	the wind farm	349.51	53	NP	place reference
1238	of small fishing vessels	357.94	53	PP-based fragment	physical entities (vessel)

1239	passenger ships	366.14	53	NP	physical entities (vessel)
1240	merchant vessels	263.18	53	NP	physical entities (vessel)
1241	the lifeboats	29.66	53	NP	physical entities (vessel)
1242	bulk carriers	739.39	53	NP	physical entities (vessel)
1243	of lifejackets	46.14	53	PP-based fragment	physical entities (equipment)
1244	the cursor	76.70	53	NP	physical entities (equipment)
1245	the dredger	65.21	53	NP	physical entities (equipment)
1246	the paper chart	91.63	53	NP	physical entities (equipment)
1247	to avoid collision	347.33	53	TO CLAUSE	notion
1248	the hazards associated	353.95	53	other NP fragment	notion
1249	control measure	482.63	53	NP	notion
1250	bridge team management	441.21	53	NP	notion
1251	was no evidence of	206.13	53	BE+ADJ./NOUN/PP	notion
1252	was no requirement for	289.92	53	BE+ADJ./NOUN/PP	notion
1253	was reported that	201.82	53	THAT CLAUSE	discourse organizer
1254	watch manager	428.49	53	NP	agent
1255	sea pilot	195.54	53	NP	agent
1256	master pilot	124.63	53	NP	agent
1257	the royal navy	100.05	53	NP	agent
1258	be secured	199.98	53	VP-P	acvity/action
1259	were fitted with	233.77	53	VP-P	acvity/action
1260	was not fitted with	314.82	53	VP-P	acvity/action
1261	pilot ordered	411.29	53	SENTENCE STEM	acvity/action
1262	master instructed	281.63	53	SENTENCE STEM	acvity/action
1263	on a heading	280.75	53	PP-based fragment	acvity/action
1264	exposure to	164.39	53	other NP fragment	acvity/action
1265	discharge of	67.66	53	NP+of fragment	acvity/action
1266	was not required	203.43	52	VP-P	stance
1267	the ability to	89.59	52	other NP fragment	stance
1268	it is apparent that	354.54	52	IT CLAUSE	stance
1269	is highly likely	498.12	52	BE+ADJ./NOUN/PP	stance
1270	impossible to	177.69	52	ADJ. fragment	stance
1271	there was no evidence of	332.21	52	SENTENCE STEM	specification of attributes
1272	the depth of	119.38	52	NP+of fragment	specification of attributes
1273	shipping notice	514.37	52	NP	regulation
1274	stability requirements	297.78	52	NP	regulation
1275	merchant shipping notice	647.06	52	NP	regulation
1276	to the hull	61.02	52	PP-based fragment	place reference
1277	its port	63.56	52	NP	place reference
1278	inland waterways	811.90	52	NP	place reference
1279	from the ship	82.74	52	PP-based fragment	physical entities (vessel)
1280	on ships	53.82	52	PP-based fragment	physical entities (vessel)
1281	in heavy weather	392.35	52	PP-based fragment	physical entities (others)



1282	breathing apparatus	796.48	52	NP	physical entities (equipment)
1283	tie bolts	784.23	52	NP	physical entities (equipment)
1284	a bilge	64.92	52	NP	physical entities (equipment)
1285	the forward mooring	273.52	52	NP	physical entities (equipment)
1286	the starboard engine	173.57	52	NP	physical entities (equipment)
1287	in restricted visibility	460.15	52	PP-based fragment	notion
1288	local control	310.41	52	NP	notion
1289	coastguard agency is recommended to	627.18	52	SENTENCE STEM	agent
1290	to mariners	143.26	52	PP-based fragment	agent
1291	relief skipper	416.83	52	NP	agent
1292	officers and crew	441.45	52	NP	agent
1293	was passed	80.89	52	VP-P	acvity/action
1294	fell overboard	511.90	52	VP-A	acvity/action
1295	to refloat	211.99	52	TO CLAUSE	acvity/action
1296	to activate	174.31	52	TO CLAUSE	acvity/action
1297	to launch	89.38	52	TO CLAUSE	acvity/action
1298	to assist in	152.14	52	TO CLAUSE	acvity/action
1299	powerboat racing	651.29	52	NP	acvity/action
1300	fishing operations	200.94	52	NP	acvity/action
1301	required to carry	365.10	51	VP-P	stance
1302	would take	199.83	51	VP-A	stance
1303	regulations require	438.86	51	SENTENCE STEM	stance
1304	vessel could	45.12	51	SENTENCE STEM	stance
1305	it is highly likely	499.32	51	IT CLAUSE	stance
1306	not comply with	292.98	51	VP-A	specification of attributes
1307	port marine safety code	599.85	51	NP	regulation
1308	the breakwater	68.75	51	NP	place reference
1309	wheelhouse roof	555.48	51	NP	place reference
1310	poop deck	503.02	51	NP	place reference
1311	these vessels	127.02	51	NP	physical entities (vessel)
1312	navigation lights	469.44	51	NP	physical entities (equipment)
1313	the hazards associated with	362.32	51	other NP fragment	notion
1314	fishing vessel certificate	446.16	51	NP	notion
1315	despite this	212.16	51	PP-based fragment	discourse organizer
1316	merchant seamen	640.12	51	NP	agent
1317	stowed on	174.13	51	VP-P	acvity/action
1318	was attached	119.26	51	VP-P	acvity/action
1319	navigating in	176.43	51	VP-A	acvity/action
1320	the master decided	202.86	51	SENTENCE STEM	acvity/action
1321	of water ingress	372.66	51	PP-based fragment	acvity/action
1322	passage planning and	320.36	51	other NP fragment	acvity/action
1323	the dredging	31.49	51	NP	acvity/action
1324	the distress	10.82	51	NP	acvity/action

1325	stability book	523.58	51	NP	physical entities (others)
1326	have enabled	311.32	50	VP-A	stance
1327	would have increased	366.87	50	VP-A	stance
1328	vessel should	31.79	50	SENTENCE STEM	stance
1329	they would have	256.31	50	SENTENCE STEM	stance
1330	appropriate action	343.19	50	NP	stance
1331	it is evident from	339.32	50	IT CLAUSE	stance
1332	is essential that	329.66	50	BE+ADJ./NOUN/PP	stance
1333	was clear of	131.46	50	BE+ADJ./NOUN/PP	stance
1334	possible to determine	493.86	50	ADJ. fragment	stance
1335	complying with	363.51	50	VP-A	specification of attributes
1336	code of safe working practices	649.42	50	NP	regulation
1337	by the port	50.38	50	PP-based fragment	place reference
1338	and engine room	255.06	50	other NP fragment	place reference
1339	machinery spaces	502.34	50	NP	place reference
1340	river runner	638.54	50	NP	physical entities (vessel)
1341	by vhf radio	157.18	50	PP-based fragment	physical entities (equipment)
1342	dredge bags	694.19	50	NP	physical entities (equipment)
1343	fork lift	692.28	50	NP	physical entities (equipment)
1344	band radar	542.58	50	NP	physical entities (equipment)
1345	automatic identification system	607.48	50	NP	physical entities (equipment)
1346	accordance with the requirements of	519.69	50	other NP fragment	notion
1347	intact stability	540.83	50	NP	notion
1348	marine environment	430.77	50	NP	notion
1349	operational procedures	370.45	50	NP	notion
1350	the control measures	270.14	50	NP	notion
1351	a risk of collision	431.01	50	NP	notion
1352	report concluded that	378.71	50	SENTENCE STEM	discourse organizer
1353	in the mail	55.25	50	PP-based fragment	discourse organizer
1354	it was noted that	355.55	50	IT CLAUSE	discourse organizer
1355	although not	69.56	50	ADV fragment	discourse organizer
1356	and chief engineer	236.37	50	other NP fragment	agent
1357	the navigational watch	273.18	50	NP	agent
1358	grounded on	181.79	50	VP-A	activity/action
1359	to abort	167.99	50	TO CLAUSE	activity/action
1360	the pilot ordered	276.80	50	SENTENCE STEM	activity/action
1361	the master ordered	202.07	50	SENTENCE STEM	activity/action
1362	collision between	249.37	50	other NP fragment	activity/action
1363	air supply	477.87	50	NP	activity/action
1364	cargo discharge	355.27	50	NP	activity/action
1365	gale force	621.80	50	NP	physical entities (others)
1366	after the collision	334.81	49	PP-based fragment	time reference
1367	is intended to	198.72	49	VP-P	stance

1368	should ensure that	270.69	49	THAT CLAUSE	stance
1369	normal practice	429.49	49	NP	stance
1370	as necessary	165.24	49	ADV fragment	stance
1371	of the forward	34.64	49	PP-based fragment	place reference
1372	in the galley	182.24	49	PP-based fragment	place reference
1373	the forward end of	150.27	49	NP+of fragment	place reference
1374	forward end of	186.03	49	NP+of fragment	place reference
1375	cargo holds	443.08	49	NP	place reference
1376	control station	352.33	49	NP	place reference
1377	the pool	45.14	49	NP	place reference
1378	the dock	11.72	49	NP	place reference
1379	the harbour entrance	331.98	49	NP	place reference
1380	the wheelhouse roof	325.32	49	NP	place reference
1381	cargo vessels	102.50	49	NP	physical entities (vessel)
1382	vhf radios	563.44	49	NP	physical entities (equipment)
1383	the conveyor	44.72	49	NP	physical entities (equipment)
1384	a radar	40.85	49	NP	physical entities (equipment)
1385	the pots	25.06	49	NP	physical entities (equipment)
1386	the steering gear	282.58	49	NP	physical entities (equipment)
1387	equipment on board	280.65	49	NP	physical entities (equipment)
1388	the general alarm	91.60	49	NP	physical entities (equipment)
1389	meet the requirements of	428.57	49	VP-A	notion
1390	significant wave height	620.68	49	NP	notion
1391	estimated that	196.23	49	THAT CLAUSE	discourse organizer
1392	it was reported that	337.11	49	IT CLAUSE	discourse organizer
1393	therefore not	103.74	49	ADV fragment	discourse organizer
1394	for merchant seamen	567.41	49	PP-based fragment	agent
1395	and the mate	42.90	49	other NP fragment	agent
1396	crews of	32.17	49	NP+of fragment	agent
1397	the junior deckhand	373.27	49	NP	agent
1398	class pilot	316.27	49	NP	agent
1399	the cadet	20.65	49	NP	agent
1400	safe navigational watch	517.31	49	NP	agent
1401	forth guardsman	838.80	49	NP	agent
1402	posed by	322.09	49	VP-P	acvity/action
1403	to board	81.62	49	TO CLAUSE	acvity/action
1404	of grounding	27.78	49	PP-based fragment	acvity/action
1405	the salvage	18.20	49	NP	acvity/action
1406	the mayday	30.81	49	NP	physical entities (others)
1407	were required to be	256.29	48	VP-P	stance
1408	recommendations shall	466.56	48	SENTENCE STEM	stance
1409	the crew would	93.41	48	SENTENCE STEM	stance
1410	his ability to	249.91	48	other NP fragment	stance

1411	was possible	25.89	48	BE+ADJ./NOUN/PP	stance
1412	was likely to	147.19	48	BE+ADJ./NOUN/PP	stance
1413	if necessary	321.96	48	ADV fragment	stance
1414	on a heading of	225.68	48	PP-based fragment	specification of attributes
1415	the code of practice for	406.76	48	other NP fragment	regulation
1416	loading manual	361.04	48	NP	regulation
1417	engine control room	460.04	48	NP	place reference
1418	on vessel	105.98	48	PP-based fragment	physical entities (vessel)
1419	of the vehicle	95.63	48	PP-based fragment	physical entities (vessel)
1420	passenger ship	214.14	48	NP	physical entities (vessel)
1421	anchor chain	459.46	48	NP	physical entities (equipment)
1422	starboard anchor	297.25	48	NP	physical entities (equipment)
1423	hand held vhf	596.38	48	NP	physical entities (equipment)
1424	emergency fire pump	544.77	48	NP	physical entities (equipment)
1425	good visibility	416.66	48	NP	notion
1426	marine accident investigation	556.05	48	NP	notion
1427	irrespective of	201.78	48	ADJ. fragment	discourse organizer
1428	was loaded	103.42	48	VP-P	acvity/action
1429	have alerted	276.09	48	VP-A	acvity/action
1430	berth in	56.18	48	VP-A	acvity/action
1431	the master instructed	185.54	48	VP-A	acvity/action
1432	master informed	219.96	48	SENTENCE STEM	acvity/action
1433	of pilotage	39.37	48	PP-based fragment	acvity/action
1434	scallop dredging	647.96	48	NP	acvity/action
1435	the sinking	51.28	48	NP	acvity/action
1436	been required	41.02	47	VP-P	stance
1437	appeared to have	276.40	47	VP-A	stance
1438	require that	139.43	47	THAT CLAUSE	stance
1439	the skipper would	93.74	47	SENTENCE STEM	stance
1440	it was safe	335.53	47	IT CLAUSE	stance
1441	it was evident	332.60	47	IT CLAUSE	stance
1442	it is highly likely that	431.27	47	IT CLAUSE	stance
1443	it is essential that	325.88	47	IT CLAUSE	stance
1444	is responsible for	271.23	47	BE+ADJ./NOUN/PP	stance
1445	would not be	209.84	47	BE+ADJ./NOUN/PP	stance
1446	also possible	191.01	47	ADJ. fragment	stance
1447	the result of	84.27	47	NP+of fragment	specification of attributes
1448	management certificate	286.60	47	NP	regulation
1449	within the port	213.59	47	PP-based fragment	place reference
1450	of its vessels	96.59	47	PP-based fragment	physical entities (vessel)
1451	a fleet of	180.74	47	NP+of fragment	physical entities (vessel)
1452	city cruises	670.79	47	NP	physical entities (vessel)
1453	ski boat	457.30	47	NP	physical entities (vessel)

1454	sister vessel	290.12	47	NP	physical entities (vessel)
1455	the speedboat	74.31	47	NP	physical entities (vessel)
1456	the liferafts	15.74	47	NP	physical entities (vessel)
1457	lift truck	601.61	47	NP	physical entities (equipment)
1458	the switchboard	48.55	47	NP	physical entities (equipment)
1459	ship safety	60.87	47	NP	notion
1460	passage plans	434.99	47	NP	notion
1461	dead slow	654.93	47	ADJ. fragment	notion
1462	was caused by	233.95	47	VP-P	discourse organizer
1463	found to be	144.81	47	VP-P	discourse organizer
1464	responded to	126.01	47	VP-A	discourse organizer
1465	recommends that	272.09	47	THAT CLAUSE	discourse organizer
1466	were unaware of	223.00	47	BE+ADJ./NOUN/PP	discourse organizer
1467	second officer and	210.46	47	other NP fragment	agent
1468	remaining crew	274.74	47	NP	agent
1469	powered by	260.80	47	VP-P	acvity/action
1470	altered course to	385.00	47	VP-A	acvity/action
1471	to heave	179.53	47	TO CLAUSE	acvity/action
1472	to abandon	123.84	47	TO CLAUSE	acvity/action
1473	for passage	55.32	47	PP-based fragment	acvity/action
1474	deck wash	395.70	47	NP	acvity/action
1475	cargo loading	260.23	47	NP	acvity/action
1476	survey and inspection	520.25	47	NP	acvity/action
1477	was required to be	188.26	46	VP-P	stance
1478	ability of	49.57	46	NP+of fragment	stance
1479	usual practice	501.71	46	NP	stance
1480	was safe to	175.21	46	BE+ADJ./NOUN/PP	stance
1481	fully aware	393.34	46	ADJ. fragment	stance
1482	flow of	95.14	46	NP+of fragment	specification of attributes
1483	over the stern	329.89	46	PP-based fragment	place reference
1484	of a ship	84.61	46	PP-based fragment	physical entities (vessel)
1485	passenger ferry	369.50	46	NP	physical entities (vessel)
1486	cargo tank	253.67	46	NP	physical entities (vessel)
1487	admiral blake	817.28	46	NP	physical entities (equipment)
1488	bilge pumping	520.52	46	NP	physical entities (equipment)
1489	oil heater	470.80	46	NP	physical entities (equipment)
1490	hydraulic system	274.21	46	NP	physical entities (equipment)
1491	navigation equipment	247.06	46	NP	physical entities (equipment)
1492	the bulwarks	61.69	46	NP	physical entities (equipment)
1493	the bilge alarm	236.51	46	NP	physical entities (equipment)
1494	fork lift truck	780.15	46	NP	physical entities (equipment)
1495	thermal oil heater	692.74	46	NP	physical entities (equipment)
1496	of work and rest	171.52	46	PP-based fragment	notion

1497	primary means of navigation	599.83	46	NP	notion
1498	stability awareness	359.04	46	NP	notion
1499	manual control	242.92	46	NP	notion
1500	served as	253.32	46	VP-P	discourse organizer
1501	was connected to	154.58	46	VP-P	discourse organizer
1502	acted as	292.81	46	VP-A	discourse organizer
1503	acting as	261.14	46	VP-A	discourse organizer
1504	provided that	24.25	46	THAT CLAUSE	discourse organizer
1505	figure shows	346.48	46	SENTENCE STEM	discourse organizer
1506	it was found	197.60	46	IT CLAUSE	discourse organizer
1507	is at figure	262.95	46	BE+ADJ./NOUN/PP	discourse organizer
1508	for the skipper	15.49	46	PP-based fragment	agent
1509	the mate and	48.52	46	other NP fragment	agent
1510	fish industry	400.63	46	NP	agent
1511	two crew	83.25	46	NP	agent
1512	fish industry authority	602.18	46	NP	agent
1513	sea fish industry	537.32	46	NP	agent
1514	dragged overboard	525.89	46	VP-P	acvitivity/action
1515	secured in	72.35	46	VP-P	acvitivity/action
1516	were stowed	207.27	46	VP-P	acvitivity/action
1517	alerted to	80.44	46	VP-A	acvitivity/action
1518	wearing a pfd	617.11	46	VP-A	acvitivity/action
1519	had worked on board	336.24	46	VP-A	acvitivity/action
1520	off the berth	422.26	46	PP-based fragment	acvitivity/action
1521	would have provided	273.19	45	VP-A	stance
1522	inability to	175.75	45	other NP fragment	stance
1523	are likely	182.11	45	BE+ADJ./NOUN/PP	stance
1524	be effective	134.12	45	BE+ADJ./NOUN/PP	stance
1525	vessels code	110.50	45	NP	regulation
1526	to full astern	325.54	45	PP-based fragment	place reference
1527	the rear	76.78	45	NP	place reference
1528	on the bridge and	91.89	45	PP-based fragment	place reference
1529	of the wreck	121.48	45	PP-based fragment	place reference
1530	astern of	16.10	45	NP+of fragment	place reference
1531	the hatch cover	286.30	45	NP	place reference
1532	the aft compartment	87.16	45	NP	place reference
1533	manning levels	494.92	45	NP	physicla entities (others)
1534	each vessel	67.69	45	NP	physical entities (vessel)
1535	mooring line	320.62	45	NP	physical entities (equipment)
1536	a mooring	33.66	45	NP	physical entities (equipment)
1537	the propeller pitch	268.74	45	NP	physical entities (equipment)
1538	the risk of falling	345.33	45	NP	notion
1539	be used as	222.92	45	VP-P	discourse organizer



1540	was reported to be	258.83	45	VP-P	discourse organizer
1541	was considered to	128.97	45	VP-P	discourse organizer
1542	also stated that	271.81	45	VP-A	discourse organizer
1543	in such circumstances	239.43	45	PP-based fragment	discourse organizer
1544	although it is	293.44	45	ADV fragment	discourse organizer
1545	by the bridge	43.27	45	PP-based fragment	agent
1546	bridge team and	215.01	45	other NP fragment	agent
1547	the bosun and	73.78	45	other NP fragment	agent
1548	duty officer	238.94	45	NP	agent
1549	ship owners	225.42	45	NP	agent
1550	a man overboard	153.55	45	NP	agent
1551	the sea pilot	87.16	45	NP	agent
1552	struck by	236.03	45	VP-P	acvitivity/action
1553	be released	199.65	45	VP-P	acvitivity/action
1554	was pulled	142.82	45	VP-P	acvitivity/action
1555	assigned to	125.48	45	VP-P	acvitivity/action
1556	was approaching	139.29	45	VP-A	acvitivity/action
1557	heading to	19.91	45	VP-A	acvitivity/action
1558	to discharge	45.57	45	TO CLAUSE	acvitivity/action
1559	over the watch	331.71	45	PP-based fragment	acvitivity/action
1560	in consultation with	276.94	45	PP-based fragment	acvitivity/action
1561	execution of	163.48	45	NP+of fragment	acvitivity/action
1562	voyage planning	427.59	45	NP	acvitivity/action
1563	emergency preparedness	488.92	45	NP	acvitivity/action
1564	collisions at sea	460.92	45	NP	acvitivity/action
1565	was aground	153.60	45	BE+ADJ./NOUN/PP	acvitivity/action
1566	intends to	181.35	44	VP-P	stance
1567	planned to	26.78	44	VP-A	stance
1568	to require	40.35	44	TO CLAUSE	stance
1569	it would not	168.89	44	IT CLAUSE	stance
1570	be aware	133.83	44	BE+ADJ./NOUN/PP	stance
1571	and possibly	81.91	44	ADV fragment	stance
1572	an angle of	226.70	44	NP+of fragment	specification of attributes
1573	vessels code of	213.22	44	NP+of fragment	regulation
1574	safety of life at sea	432.14	44	NP	regulation
1575	stcw ii certificate	531.21	44	NP	regulation
1576	fishing vessels code	327.02	44	NP	regulation
1577	container terminal	467.72	44	NP	place reference
1578	the poop deck	260.07	44	NP	place reference
1579	for ships	60.18	44	PP-based fragment	physical entities (vessel)
1580	for a vessel	98.57	44	PP-based fragment	physical entities (vessel)
1581	all ships	176.89	44	NP	physical entities (vessel)
1582	pilot boat	171.43	44	NP	physical entities (vessel)

1583	of the winch	46.34	44	PP-based fragment	physical entities (equipment)
1584	wing console	521.77	44	NP	physical entities (equipment)
1585	navigational aids	472.66	44	NP	physical entities (equipment)
1586	trawl wire	447.37	44	NP	physical entities (equipment)
1587	high voltage	440.97	44	NP	physical entities (equipment)
1588	stability information	238.91	44	NP	notion
1589	pull of	73.55	44	NP+of fragment	notion
1590	sailing directions	552.31	44	NP	notion
1591	contract on board	303.57	44	NP	notion
1592	laid out	347.20	44	VP-P	discourse organizer
1593	shown at figure	548.79	44	VP-P	discourse organizer
1594	made aware of	242.67	44	VP-A	discourse organizer
1595	was contrary to	168.44	44	BE+ADJ./NOUN/PP	discourse organizer
1596	was not aware of	182.01	44	BE+ADJ./NOUN/PP	discourse organizer
1597	by the chief officer	214.09	44	PP-based fragment	agent
1598	by the coastguard	149.01	44	PP-based fragment	agent
1599	with the skipper	24.23	44	PP-based fragment	agent
1600	the chief engineer and	156.19	44	other NP fragment	agent
1601	two crewmen	354.65	44	NP	agent
1602	the master and pilot	238.46	44	NP	agent
1603	the port captain	222.99	44	NP	agent
1604	contained within	307.28	44	VP-P	acvity/action
1605	generated by	247.88	44	VP-P	acvity/action
1606	was manoeuvred	130.56	44	VP-P	acvity/action
1607	is maintained	172.55	44	VP-P	acvity/action
1608	towing operations	298.89	44	NP	acvity/action
1609	the stowage	7.63	44	NP	acvity/action
1610	was on board	14.18	44	BE+ADJ./NOUN/PP	acvity/action
1611	master decided to	212.92	44	SENTENCE STEM	activity/action
1612	board at the time of	330.99	43	VP-A	time reference
1613	he intended to	192.41	43	SENTENCE STEM	stance
1614	the vessel could	66.64	43	SENTENCE STEM	stance
1615	was capable	109.93	43	BE+ADJ./NOUN/PP	stance
1616	is no requirement	261.83	43	BE+ADJ./NOUN/PP	stance
1617	are likely to	216.57	43	BE+ADJ./NOUN/PP	stance
1618	was capable of	147.18	43	BE+ADJ./NOUN/PP	stance
1619	used on board	229.96	43	VP-P	specification of attributes
1620	bight of	158.12	43	NP+of fragment	place reference
1621	close quarters	519.45	43	NP	place reference
1622	wheelhouse door	314.74	43	NP	place reference
1623	a harbour	15.57	43	NP	place reference
1624	on a vessel	66.96	43	PP-based fragment	physical entities (vessel)
1625	for the ship	23.48	43	PP-based fragment	physical entities (vessel)

1626	of the vessel was	8.74	43	PP-based fragment	physical entities (vessel)
1627	seatruck ferries	577.51	43	NP	physical entities (vessel)
1628	any vessel	27.71	43	NP	physical entities (vessel)
1629	the rudders	58.48	43	NP	physical entities (equipment)
1630	hatch lid	524.25	43	NP	physical entities (equipment)
1631	drive shaft	492.64	43	NP	physical entities (equipment)
1632	a lifebuoy	181.28	43	NP	physical entities (equipment)
1633	the stopper	61.16	43	NP	physical entities (equipment)
1634	the hoist	52.36	43	NP	physical entities (equipment)
1635	the capacitor	51.79	43	NP	physical entities (equipment)
1636	the rigging	47.47	43	NP	physical entities (equipment)
1637	the engine telegraph	247.40	43	NP	physical entities (equipment)
1638	navigational hazards	379.02	43	NP	notion
1639	deficiencies identified	329.95	43	NP	notion
1640	referred to	122.35	43	VP-P	discourse organizer
1641	there is no evidence to suggest	280.50	43	SENTENCE STEM	discourse organizer
1642	and consequently	37.68	43	ADV fragment	discourse organizer
1643	from the master	15.01	43	PP-based fragment	agent
1644	marine accident investigation branch	658.01	43	NP	agent
1645	marine office	334.13	43	NP	agent
1646	the deck crew	109.11	43	NP	agent
1647	been fitted with	193.35	43	VP-P	acvity/action
1648	turn to port	340.96	43	VP-A	acvity/action
1649	informed the master	236.51	43	VP-A	acvity/action
1650	of capsiz	65.44	43	PP-based fragment	acvity/action
1651	for the passage	137.25	43	PP-based fragment	acvity/action
1652	the collision and	45.41	43	other NP fragment	acvity/action
1653	traffic separation	560.73	43	NP	acvity/action
1654	the hoisting	44.48	43	NP	acvity/action
1655	dry powder	614.74	43	NP	physical entities (others)
1656	the ability of	104.34	42	NP+of fragment	stance
1657	common practice	386.62	42	NP	stance
1658	is also possible	317.95	42	BE+ADJ./NOUN/PP	stance
1659	is not possible	232.16	42	BE+ADJ./NOUN/PP	stance
1660	did not comply with	361.05	42	VP-A	specification of attributes
1661	surface of	44.37	42	NP+of fragment	specification of attributes
1662	causes of	64.36	42	NP+of fragment	specification of attributes
1663	vessel code	44.71	42	NP	regulation
1664	hatch coaming	513.97	42	NP	place reference
1665	in the boat	24.93	42	PP-based fragment	physical entities (vessel)
1666	banana boat	433.03	42	NP	physical entities (vessel)
1667	the messenger line	289.71	42	NP	physical entities (vessel)
1668	the other vessel	72.30	42	NP	physical entities (vessel)

1669	whipping drum	555.94	42	NP	physical entities (equipment)
1670	lift car	512.02	42	NP	physical entities (equipment)
1671	lifting appliances	471.70	42	NP	physical entities (equipment)
1672	port anchor	187.56	42	NP	physical entities (equipment)
1673	the hooks	37.43	42	NP	physical entities (equipment)
1674	the lift car	334.84	42	NP	physical entities (equipment)
1675	the propeller shaft	274.27	42	NP	physical entities (equipment)
1676	the auxiliary engine	226.13	42	NP	physical entities (equipment)
1677	the starboard anchor	196.21	42	NP	physical entities (equipment)
1678	exhaust system	277.19	42	NP	physical entities (equipment)
1679	of the risk of	79.89	42	PP-based fragment	notion
1680	dangers associated with	382.68	42	other NP fragment	notion
1681	stability condition	275.96	42	NP	notion
1682	annual self certification	588.94	42	NP	notion
1683	speed over the ground	643.11	42	NP	notion
1684	maritime safety	155.48	42	NP	notion
1685	reported as	99.39	42	VP-P	discourse organizer
1686	was limited to	155.45	42	VP-P	discourse organizer
1687	have contributed to	218.83	42	VP-A	discourse organizer
1688	to respond to	145.62	42	TO CLAUSE	discourse organizer
1689	and subsequently	57.47	42	ADV fragment	discourse organizer
1690	and its crew	122.21	42	other NP fragment	agent
1691	fishing vessel owners	350.18	42	NP	agent
1692	the watch oow	238.39	42	NP	agent
1693	sea fish industry authority	567.20	42	NP	agent
1694	none of the crew	336.18	42	NP	agent
1695	bridge watchkeepers	334.13	42	NP	agent
1696	watch oow	283.87	42	NP	agent
1697	been secured	152.48	42	VP-P	acvity/action
1698	was manufactured	142.63	42	VP-P	acvity/action
1699	maintained at	140.09	42	VP-P	acvity/action
1700	positioned on	132.59	42	VP-P	acvity/action
1701	board vessels	43.66	42	VP-A	acvity/action
1702	alter course to	352.25	42	VP-A	acvity/action
1703	sailed on	122.41	42	VP-A	acvity/action
1704	to deploy	163.35	42	TO CLAUSE	acvity/action
1705	and maintenance of	110.45	42	NP+of fragment	acvity/action
1706	her berth	271.64	42	NP	acvity/action
1707	the master informed	144.62	42	SENTENCE STEM	activity/action
1708	following the collision	280.70	41	VP-A	time reference
1709	were intended	87.21	41	VP-P	stance
1710	required to comply with	316.15	41	VP-P	stance
1711	required to have	116.72	41	VP-P	stance

1712	bold venture	713.57	41	NP	stance
1713	to full ahead	284.41	41	PP-based fragment	place reference
1714	on either side	197.20	41	PP-based fragment	place reference
1715	the astern	15.80	41	NP	place reference
1716	on the bridge at	150.91	41	PP-based fragment	place reference
1717	to the berth	61.26	41	PP-based fragment	place reference
1718	the port bridge wing	337.85	41	NP	place reference
1719	upturned hull	483.07	41	NP	place reference
1720	precautionary area	426.24	41	NP	place reference
1721	the walkway	69.37	41	NP	place reference
1722	the working deck	192.77	41	NP	place reference
1723	a vessel of	99.76	41	NP+of fragment	physical entities (vessel)
1724	immersion suit	586.99	41	NP	physical entities (equipment)
1725	centre console	427.99	41	NP	physical entities (equipment)
1726	ecdis display	375.42	41	NP	physical entities (equipment)
1727	detection system	314.70	41	NP	physical entities (equipment)
1728	service pump	265.65	41	NP	physical entities (equipment)
1729	the bilge pump	212.81	41	NP	physical entities (equipment)
1730	there is no requirement	401.15	41	SENTENCE STEM	notion
1731	heel test	448.82	41	NP	notion
1732	structural failure	394.05	41	NP	notion
1733	tug assistance	323.41	41	NP	notion
1734	the risk of collision	244.32	41	NP	notion
1735	demonstrate that	197.64	41	THAT CLAUSE	discourse organizer
1736	were aware of	144.38	41	BE+ADJ./NOUN/PP	discourse organizer
1737	to the chief officer	171.50	41	PP-based fragment	agent
1738	by the maritime and coastguard	419.25	41	PP-based fragment	agent
1739	of fishermen	15.84	41	PP-based fragment	agent
1740	its crews	184.92	41	NP	agent
1741	port authorities	238.91	41	NP	agent
1742	ship manager	224.27	41	NP	agent
1743	department for transport	611.21	41	NP	agent
1744	vessel traffic services	405.97	41	NP	agent
1745	ordination centre	485.04	41	NP	agent
1746	being dragged	367.75	41	VP-P	acvity/action
1747	was lowered	123.49	41	VP-P	acvity/action
1748	was manned by	240.06	41	VP-P	acvity/action
1749	passing through	324.45	41	VP-A	acvity/action
1750	for the safe operation of	302.80	41	PP-based fragment	acvity/action
1751	the installation of	99.05	41	NP+of fragment	acvity/action
1752	bilge suction	454.50	41	NP	acvity/action
1753	were on board	6.61	41	BE+ADJ./NOUN/PP	acvity/action
1754	mayday relay	617.89	41	NP	physical entities (others)

1755	admiralty chart	466.07	41	NP	physical entities (others)
1756	renewal survey	452.39	41	NP	physical entities (others)
1757	the marine accident	172.25	41	NP	physical entities (others)
1758	after the grounding	290.11	40	PP-based fragment	time reference
1759	also possible that	246.39	40	THAT CLAUSE	stance
1760	sms required	182.60	40	SENTENCE STEM	stance
1761	crew should	54.50	40	SENTENCE STEM	stance
1762	it was possible	300.08	40	IT CLAUSE	stance
1763	it was clear	240.47	40	IT CLAUSE	stance
1764	it was safe to	229.95	40	IT CLAUSE	stance
1765	was possibly	106.86	40	BE+ADJ./NOUN/PP	stance
1766	was unlikely	96.03	40	BE+ADJ./NOUN/PP	stance
1767	been possible to	170.10	40	BE+ADJ./NOUN/PP	stance
1768	more likely to	216.53	40	ADJ. fragment	stance
1769	a crew of	95.92	40	NP+of fragment	specification of attributes
1770	the international regulations	198.11	40	NP	regulation
1771	in the port of	97.08	40	PP-based fragment	place reference
1772	ahead and astern	470.91	40	ADJ. fragment	place reference
1773	in his cabin	160.26	40	PP-based fragment	place reference
1774	to the scene	88.24	40	PP-based fragment	place reference
1775	in the deck	21.45	40	PP-based fragment	place reference
1776	port boiler room	398.54	40	NP	place reference
1777	their cabins	352.82	40	NP	place reference
1778	cabin space	341.84	40	NP	place reference
1779	on board vessels	139.42	40	PP-based fragment	physical entities (vessel)
1780	national lifeboat	360.30	40	NP	physical entities (vessel)
1781	vessels engaged	206.58	40	NP	physical entities (vessel)
1782	class vessels	168.80	40	NP	physical entities (vessel)
1783	royal national lifeboat	587.16	40	NP	physical entities (vessel)
1784	small fishing vessel	234.38	40	NP	physical entities (vessel)
1785	of carbon monoxide	393.16	40	PP-based fragment	physical entities (equipment)
1786	bowsing tackle	651.47	40	NP	physical entities (equipment)
1787	floor plates	619.51	40	NP	physical entities (equipment)
1788	vertical ladder	463.95	40	NP	physical entities (equipment)
1789	trawl winch	370.99	40	NP	physical entities (equipment)
1790	control levers	359.51	40	NP	physical entities (equipment)
1791	positioning system	331.82	40	NP	physical entities (equipment)
1792	port boiler	331.18	40	NP	physical entities (equipment)
1793	winch control	209.23	40	NP	physical entities (equipment)
1794	the burner	75.57	40	NP	physical entities (equipment)
1795	radio call	308.16	40	NP	physical entities (equipment)
1796	cpp control	273.24	40	NP	physical entities (equipment)
1797	the dredge gear	248.54	40	NP	physical entities (equipment)



1798	appreciation of	161.84	40	NP+of fragment	notion
1799	a course of	125.60	40	NP+of fragment	notion
1800	bulwark height	422.81	40	NP	notion
1801	onboard procedures	315.63	40	NP	notion
1802	operating conditions	223.20	40	NP	notion
1803	navigational safety	155.92	40	NP	notion
1804	laid down	412.28	40	VP-P	discourse organizer
1805	was attached to	128.56	40	VP-P	discourse organizer
1806	was considered to be	211.45	40	VP-P	discourse organizer
1807	not considered to be	244.91	40	VP-P	discourse organizer
1808	demonstrates that	205.19	40	THAT CLAUSE	discourse organizer
1809	informed that	84.84	40	THAT CLAUSE	discourse organizer
1810	for seafarers	187.96	40	PP-based fragment	agent
1811	by maib	62.80	40	PP-based fragment	agent
1812	to masters	10.22	40	PP-based fragment	agent
1813	by the manufacturer	198.01	40	PP-based fragment	agent
1814	between the master	84.94	40	PP-based fragment	agent
1815	the international chamber of	260.55	40	NP+of fragment	agent
1816	watch officer	146.53	40	NP	agent
1817	passengers on board	250.71	40	NP	agent
1818	second bosun	311.31	40	NP	agent
1819	separated from	276.71	40	VP-P	acvity/action
1820	arrived on board	215.80	40	VP-A	acvity/action
1821	to withstand	157.38	40	TO CLAUSE	acvity/action
1822	of propulsion	24.29	40	PP-based fragment	acvity/action
1823	passage through	200.22	40	other NP fragment	acvity/action
1824	passage in	8.32	40	other NP fragment	acvity/action
1825	shipboard operations	386.44	40	NP	acvity/action
1826	tidal conditions	290.83	40	NP	physical entities (others)