https://doi.org/10.24303/lakdoi.2021.29.3.105

Keyword Analyses of a 19th Century English Expedition Journal Corpus: Focusing on Vancouver and Broughton's Expedition Journal*

Sunghwa Lee, Se-Eun Jhang** & Qi Yang
(Korea International University in Ferghana, Korea Maritime and Ocean University & Dalian Maritime University/Korea Maritime and Ocean University)

Lee, Sunghwa; Jhang, Se-Eun & Yang, Qi. (2021). Keyword analyses of a 19th century English expedition journal corpus: Focusing on Vancouver and Broughton's expedition journal. *The Linguistic Association of Korea Journal, 29*(3), 105-128. The purpose of the study is to investigate the lexical characteristics of Maritime English as used in the 18th and 19th centuries through two approaches to keyword analysis: corpus frequency-based keyword analysis and text dispersion-based keyword analysis. The target corpus used in this study is an English journal corpus consisting of Captain Vancouver and Lieutenant Broughton's expedition journal written in the late 18th and early 19th centuries. We focus on comparing important maritime-related words and essential theme-specific words used in the target corpus with those used in the BE06 corpus, representing contemporary general British English, and those used in the Contemporary Maritime English Corpus (CMEC). Through a bilateral cross-comparison of the two keyword analyses, we conclude that dispersion-based keyword analysis shows better results than frequency-based keyword analysis in that the former is more effective than the latter regarding the two criteria proposed by Egbert and Biber (2019): the content-distinctiveness of maritime-related keywords and the content-generalisability of theme-specific keywords. We also discuss some exciting findings about maritime-related keywords under a multilateral comparison of those

* An earlier version of this paper was presented at the 2021 KASELL Spring Conference on English Linguistics on June 5th, 2021, hosted by The Korean Association for the Study of English Language and Linguistics.

** Sunghwa Lee is the first author and Se-Eun Jhang, the corresponding author.

used in the target corpus and the BE06 corpus as well as in the target corpus and the CMEC.

Key Words: keyword analysis, expedition journal, frequency, text dispersion, Maritime English

1. Introduction

A logbook is a record of important events in a ship's management, operation, and navigation, following Wikipedia definition (https://en.wikipedia.org/wiki/Logbook). The "logbook" was essential to traditional navigation, and captains or commanders of a ship must concisely write operational data related to a ship such as weather conditions, directions and times of recurring events and expeditions, significant incidents, what ports were docked at when, etc., at least daily. Alongside such logs officers were expected to keep a personal journal (Allen, 2002) which incorporated elements from the log alongside their personal observations and thoughts or discussions. The present target corpus is a published book from such a journal, entitled "A narrative or journal of a voyage of discovery to the North Pacific Ocean and round the world performed in the years 1791, 1792, 1793, 1794 and 1795." Such a journal mainly contains maritime-related words and theme-specific content words, which are significantly different from a diary written in general English, as well as function words and general content words, which are similar to an ordinary diary. The present sudy focuses on maritime-related words and theme-specific content words that maritime expedition journals written in the 18th and 19th centuries usually contain. Thus the present study also discusses how we obtain these specialized words to access the textual features which are domain-specific.

Professionals in any domain-specific area have always been concerned with the specific vocabulary of their field of specialized knowledge. However, it has been suggested that specialized languages need to be studied by corpus linguists interested in English for Specific Purposes (ESP) with the help of specialists in a given field, whenever necessary (Chung & Nation, 2004; Gabrielatos & Sarmento, 2006; Hong & Jhang, 2010; Hou, 2014; Liu, 2021; Srichai, 2017; Qi, 2019; Xu, 2021). Most recently, Xu (2021) developed a list of specialized vocabulary (technical vocabulary in her term) in Maritime English, as an example of a domain-specific field, by using a hybrid method of the combination of

keyword analysis and specialized vocabulary rating scale, as "[K]eyword analysis is used to identify the words that are especially characteristic of the texts in a target discourse domain" (Gries, 2021; Scott, 1997). It has been discussed that a list of keywords carries significant implications of the text, facilitates understanding of the main point of the text, and reflects essential themes that characterize what the text is about (Baker, 2004; Bondi, 2010; Scott & Tribble, 2006; Stubbs, 2010).

It is well-known in corpus linguistics that the choice of a reference corpus can cause a list of the keywords generated to differ (Gelus & Hirch, 2019; Goh, 2011). What is more important than the choice of a reference corpus, however, is to consider which approach to keyword analysis should be taken in identifying the most appropriate keywords for the target corpus. This critical reason is that the keywords generated may be rather different even though we conducted our comparison of the target corpus against the same reference corpus. Other than Gries' (2021) recent research on the hybrid of frequency and text dispersion keyness, the most commonly used method of calculating keyness in corpus linguistics is frequency-based keyword analysis. After Egbert and Biber (2019) proposed a new approach to keyword analysis called a text dispersion-based keyword analysis, this new analytic method has been included in Wordsmith Tools 8.0 (Scott, 2020).

Considering these recent advances, it is wise to experiment by identifying a list of keywords generated from a domain-specific field such as a nautical expedition through comparison against a large general corpus and analyze the keywords through several different approaches to keyword analysis. Due to the space limits here, the present paper will focus on a cross-comparison of the keywords generated using the two different approaches to keyword analysis: frequency-based and text dispersion-based keyword analyses. The purpose of the study is to investigate lexical characteristics of Maritime English used in the 18th and 19th centuries through the two approaches to keyword analyses: corpus frequency-based keyword analysis and text dispersion-based keyword analysis. The corpus used in the present study is an English expedition journal corpus that consists of Vancouver and Broughton's expedition journal narrative written in the late 18th and early 19th centuries. We focus on a cross-comparison of maritime-related vital words used in the present target corpus with those used in the BE06 corpus[1)] as contemporary General British English as well as those used in the Contemporary Maritime English Corpus (CMEC).

In the present paper, we try to answer the following research questions.

(1) What are the lexical characteristics of Maritime English used in the

1) See subsection 3.1 for more detailed information about the BE06 corpus.

18th and 19th centuries? If keywords carry significant implications of a text, facilitate understanding of the main point of the text, and reflect important themes that characterize what the text is about, which approach to keyword analysis shows "a most welcome keyword list" to access the textual features which are domain-specific?

(2) Have these 18th and 19th century maritime-related words and usages been retained in the contemporary English, as represented in the BE06 corpus as general English as well as in the CMEC as ESP?

2. Previous Studies on Keyword Analysis

2.1. Traditional Approach: Corpus Frequency Keywords

It is well-known in corpus linguistics that keyword analysis is one of the most widely used methods to analyze the essential words that are domain-specific as compared against an equivalent or larger general corpus based on frequency (Baker, 2004; Baron, Rayson, & Archer, 2009; Gabrelatos, 2018; Scott, 1997, 2016, 2020).

According to Scott (1997), a keyword is defined as a word that occurs with unusual frequency in a given text. This does not mean only high frequency, but the unusual frequency, through comparison with a reference corpus of some kind. As Baker (2004) discussed, keywords appear to be more frequent or infrequent in a particular text or corpus than in a reference. Scott and Tribble (2006) observe that keywords are derived from comparative quantitative corpus analysis. They define "keyness" as a term used in corpus linguistics to describe a word or phrase's quality of being "key" in its context. They also emphasize the choice and type of reference corpora. Keyness is calculated based on statistical measures such as log-likelihood or chi-square. According to Scott (2016, 2020), "[A] word is said to be 'key' if its frequency in the text when compared with its frequency in a reference corpus is such that the statistical probability as computed by an appropriate procedure is smaller than or equal to a p-value specified by the user." Bondi (2010) introduces the basic concept of keywords to facilitate understanding of a text's main point, constituting chains of repetition in the text or a corpus. In analyzing a list of keywords, Stubbs (2010) helps us understand that content words directly indicate the propositional content of texts, while function words are the most frequent in the language. As Scott and Tribble (2006) claimed, keywords can carry significant implications of a text

that discover the two properties: One is the aboutness and content of the text, and the other is a style of the text.

Most recently and more specifically, Liu (2021) explores the diachronic change of keyness characteristics focusing on the trends of marine environment protection based on a self-compiled corpus, compared with the BNC Baby Written Corpus. Her corpus frequency-based keyword analysis uncovers the trends more quantitatively and objectively, demonstrating the themes through unusually frequently high ranked keywords generated from the comparison with a reference corpus.

2.2. New Approaches: Text Dispersion and Hybrid of Frequency and Text Dispersion

Egbert and Biber (2019) claim that it has been proved that frequency-based calculation is probably not ideal since dispersion is not taken into consideration. They propose a text dispersion keyword analysis that can evaluate the effectiveness of a keyword analysis concerning two criteria: content-distinctiveness and content-generalisability. The former requires keywords that are genuinely typical in a specific domain and point out the textual features which are distinguished from the other domains. The latter needs keywords that have the quality to be generalized to the other texts in a similar domain and "offer insight into the actual content- 'aboutness' of those texts."

Xu and Jhang (2020) suggest combining the frequency-based method and the dispersion-based method for analysis of a specialized corpus to develop a consolidated keyword list. Xu (2021) then carries out a hybrid method of two different keyword analyses: a frequency-based keyword analysis proposed by Scott (2016) and a text dispersion-based keyword analysis proposed by Egbert and Biber (2019) to generate a list of commonly-used specialized vocabulary in Maritime English.

Most recently, Gries (2021) claims that text dispersion-based keyword analysis proposed by Egber and Biber (2019) can still be profitably extended by utilizing both frequency and dispersion for keyness computations. He proposes a new two-dimensional approach to keyness and exemplifies it on the basis of the Clinton-Trump Corpus and the British National Corpus. This present paper focuses on comparing the two approaches to keywords analysis and shows that the keywords generated are different. We thus claim that text dispersion-based keyword analysis generates "a most welcome keyword list" to access the textual features which are domain-specific.

3. Data and Methodology

3.1. Corpora

The Vancouver and Broughton Corpus (VBC) as a target corpus was created from a book by Captain George Vancouver and Lieutenant William Robert Broughton of the Royal Navy (1802), entitled *A narrative or journal of a voyage of discovery to the North Pacific Ocean and round the world performed in the years 1791, 1792, 1793, 1794 and 1795*. Captain Vancouver commanded the HMS Discovery sloop of war, and Commander Lieutenant Broughton led its armed tender HMS Chatham.

We downloaded both PDF and TXT forms of this book from the Internet Archive website at https://archive.org. We then carefully cross-checked the full-text version with the original microfilmed PDF to locate and repair any text-conversion errors. WordSmith Tools 8.0 version (Scott, 2020) was used to calculate the total data from the corrected text.

The current target corpus, now called the VBC, comprises 26,600 words in token and 3,656 words in type and 17 texts, as shown in Table 1.

Table 1. General statistical information of the VBC

Text File	File Size	Token	Type	TTR	STTR
Overall	154,864	26,602	3,656	13.98	40.84
01.txt	9,830	1,637	622	38.92	43.00
02.txt	7,652	1,388	508	37.85	38.60
03.txt	8,890	1,586	519	33.64	41.70
...	...	...	...	...	...
17.txt	11,087	1,871	646	35.01	42.70

As expected in the title of this journal, the expedition took place from 1791 to 1795. The journey started from the UK and focused on the west coast of North and South America and the Sandwich Islands (Hawai'i) via Australia and New Zealand. So we carefully divided the single book into 17 texts, each of which ranges average 1,600 words, according to the expedition's routes. A principal reason for this division is that the number of texts is significantly important in text dispersion-based keywords analysis.

Baker's (2009) BE06 Corpus of British English, used as our reference corpus, is a one million word corpus of published general written British English. It consists of 500 files of

2,000 word samples taken from 15 genres of writing. We used the BE06 corpus as a reference corpus (accessed from https://cqpweb.lancs.ac.uk/) because two British navy officers wrote the book which forms the target corpus in the present study. We also used another corpus, i.e., Lee's (2016) CMEC comprised of four million words, as a reference corpus, when necessary to be compared in the discussion of frequency change over time of interesting maritime-related words in question at the time of 18th and 19th centuries and 21st century.

3.2. Methods

We conducted a bilateral cross-comparison of two keyword lists using WordSmith Tools 8.0 through two keyword analyses; a traditional approach to corpus frequency keyword analysis (Scott, 2016) and a new approach to text dispersion keyword analysis (Egbert & Biber, 2019; Scott, 2020).

According to Bertels and Speelmam (2013: 24), log-likelihood could be an efficient statistical measure for both small and large corpora. This can be supported by other keyword-related journal articles about ESP research (Grabowsky, 2015; Ha, 2020; Jhang, Yu, & Lee, 2017). Furthermore, the likelihood ratio could lead to much better statistical results for small corpora, as proposed by Dunning (1993: 65). So we used a log-likelihood test (Dunning, 1993) to consider a significant level because it gives a better estimate of keyness, especially when contrasting long texts or a whole genre against a reference corpus (Scott, 2016). We then set a relatively low p-value threshold of 0.000001 (1 in one million) to produce more reliable data.

4. Results and Discussion

4.1. Comparison of Two Lists of Keywords

Two sets of keyword lists were extracted for cross-comparison through a heuristic approach[2]. On the one hand, a total of 101 text-dispersion (TD) keywords were obtained

2) Even though an anonymous review asked us to delete this term because it seems to be unscientific, we use this term in the present paper because this approach has been used in psychology, behavior economy, computer engineering, and other sciences. The following definition of heuristic was accessed from an economy website at https://www.investopedia.com/terms/h/heuristics.asp: "[A] heuristic, or a heuristic technique, is any approach to problem-solving that uses a practical method or various

by employing the text dispersion function. On the other hand, a total of 288 keywords were extracted as corpus frequency (CF) keywords with frequency basis: the top 101 keywords were compared with the TD keywords. Out of the 101 keywords, we found that 58 keywords are shared, as seen in Table 2 below, and 43 keywords are unshared, as shown in Table 3 below. Shared keywords shown in Table 2 and Table 3 are sorted by an alphabetical order.

Table 2. Top five and bottom two shared keywords extracted by the two keyword analyses

No	TD Ranking	Shared Keywords	CF Ranking
1	19	ANCHOR	58
2	9	ANCHORED	23
3	87	APPEARED	76
4	21	BAY	20
5	46	BOAT	87
...	..	...	..
57	72	WEATHER	28
58	37	WESTWARD	48

Table 3. Top five and bottom two unshared keywords extracted by the two keyword analyses

No	Ranking	TD Keywords	No	Ranking	CF Keywords
1	30	ADMIRALTY	1	90	ACCOMPANIED
2	36	ANCHORAGE	2	46	BOARD
3	60	APPEARANCE	3	65	BREEZE
4	59	ATTENDED	4	47	CALLED
5	64	BOISTEROUS	5	61	CANAL
..	..	...	..	..	...
42	86	WHENCE	42	9	WHICH
43	78	YAWL	43	79	WIND

shortcuts in order to produce solutions that may not be optimal but are sufficient given a limited timeframe or deadline. Heuristics methods are intended to be flexible and are used for quick decisions, especially when finding an optimal solution is either impossible or impractical and when working with complex data."

We will focus on a cross-comparative analysis with 43 unshared keywords, as shown in Table 3 above, which will be illustrated in the following subsection.

4.2. Quantitative Analysis of Keywords

In this subsection, we present four categories of the keywords and classify the keywords accordingly. The four categories that we suggest are as follows: (1) function words, such as *or, its,* and *whence*; (2) maritime-related words that depict marine-associated objects and/or status such as *vessels, boisterous*, and *eastward*; (3) theme-specific content words that indicate navigators' performance during their expedition, such as *habitations, Indians,* and *researches*; and (4) General content words. Table 4 shows the number of keywords classified into each of the four categories.

Table 4. Comparison of the two keyword lists according to category classification

Category Classification	Shared Keywords	Unshared Keywords		TOTAL (TD Keywords)
		Text Dispersion	Corpus Frequency	
Function words	2 (3.4%)	1 (2.3%)	3 (7.0%)	3 (3.0%)
Maritime-related words	33 (56.9%)	11 (25.6%)	19 (44.2%)	44 (43.6%)
Theme-specific content words	15 (25.9%)	13 (30.2%)	9 (20.9%)	28 (27.7%)
General content words	8 (13.8%)	18 (41.9%)	12 (27.9%)	26 (25.7%)
TOTAL	58 (100%)	43 (100%)	43 (100%)	101 (100%)

As shown in Table 4, maritime-related words appear the most frequently, followed by theme-specific content words. Function words appear the least, as expected. It shows that content-associated words appear more in TD than in CF; interestingly, CF contains more maritime-related words than TD does. It appears misleading to suggest that CF keywords describe Maritime English better than TD-based keywords do.

However, let us consider maritime-related keywords, as re-arranged in Table 5 below.

Table 5. Comparison of the unshared maritime-related keywords lists between TD and CF

Ranking	TD Keywords	Ranking	CF Keywords	Ranking	CF Keywords
36	ANCHORAGE	11	N	61	CANAL
38	EXTREMITY	18	POINT	65	BREEZE

Ranking	TD Keywords	Ranking	CF Keywords	Ranking	CF Keywords
53	SAILED	22	INLET	68	COVE
56	UNFAVOURABLE	25	E	70	NORTH
64	BOISTEROUS	26	LAND	79	WIND
68	RENDEZVOUS	35	S	80	CANOES
78	YAWL	36	HARBOUR	86	DIRECTION
88	CUTTER	40	COAST	97	OFFICERS
89	SOUNDINGS	46	BOARD		
90	COASTS	55	SEA		
100	PLEASANT	59	PASSAGE		

If we take a close look at the maritime-related keywords as illustrated in Table 5 above, more meaningful words appeared in the TD keywords. For instance, CF keywords contain basic vocabulary directions, including *N, E, S*, and *North*. As for weather-associated words, the CF keyword list includes *breeze* and *wind* that are obvious, whereas the TD list shows *unfavourable*, *boisterous*, and *pleasant* that refer to particular types of weather, which will be discussed in 4.3.1.5. Furthermore, in the TD keyword list, we can observe geographical features like *extremity*, types of vessels such as *yawl* and *cutter*, navigation-associated words, e.g., *rendezvous* and *sailed*, and anchor-associated words like *anchorage* and *soundings*. These keywords represent the content-distinctiveness of this target corpus because "[C]ontent-distinctiveness requires keywords that are truly typical in a specific domain and that point out the textual features, which are distinguished from the other domains" (Egbert & Biber, 2019: 79). As demonstrated in Table 5, a requirement of content-distinctiveness as one criterion of keywords leads us to conclude that text dispersion-based keyword analysis should be much preferred over corpus frequency-based keyword analysis. Below we will discuss keywords drawn by the text-dispersion method.

4.3. Discussion of Keyword Classification

4.3.1. Maritime-related English Keywords

Here we focus on content keywords generated by using text dispersion-based keyword analysis. A total of 44 types are classified into maritime words, which we grouped into seven categories according to themes. The seven sub-categories are illustrated in Table 6 below. In the following subsections, we discuss significant keywords (underlined and bolded words) from each of the five categories, i.e., geographical features,

direction, weather, anchor-associated words and measurements.

Table 6. Sub-categories of maritime-related English keywords

Categories	Tokens	Keywords
Geographical features	11	bay, cape, coasts, **extremity**, island(s), port, shore(s), rocks, rocky
Direction and/or location	8	east**ward**, north**ward**, south**ward**, west**ward**, lat, latitude, longitude, w
Types of vessels	8	boat(s), canoe, cutter, ship, vessel(s), yawl
Navigation-associated words	5	compass, rendezvous, sail, sailed, voyage
Weather	5	**boisterous**, gale, **pleasant**, unfavourable, weather
Anchor-associated words	5	anchor, anchorage, anchored, bore, **soundings**
Measurements	2	**fathoms**, leagues
	44	

4.3.1.1. Geographical Features: *Extremity*

The word *extremity*, defined by the Oxford dictionary as "the furthest point, especially from the centre", does not seem to be a familiar one, although we can easily figure out the meaning by inferring from "extreme". A total of eight tokens were found in the VBC. As seen in (1), the word refers to the furthest point of bays or sound.

(1) a. The point constituting the west **extremity** of these bays was called Point Roberts.
b. The high rocky bluff point forming the S. W. **extremity** of the sound, was distinguished by the name of Bald Head

Interestingly, no tokens were found in the BE06 corpus; it seems to be an archaic form. So in order to see the possibility that *extremity* is a maritime word rather than an archaic form, concordance was employed with the CMEC. And eight tokens were found in the CMEC. Sentence (2) from the CMEC adopts extremity when explaining the pivoting point.

(2) The pivoting point is the **extremity** of the perpendicular (steady turning radius) from the centre of the turn (the centre-point) onto the fore and aft line of the ship.

By exploring concordances of the VBC, the BE06 corpus, and the CMEC, we conclude that *extremity* is maritime vocabulary that has been being used in its domain.

4.3.1.2. Direction: *Eastward, Northward, Southward, Westward*

It is natural that direction-associated words frequently appeared, considering that the target text is about exploring the world. For example, *northward, westward, southward,* and *eastward* appeared 16, 15, 12, and 11 times, respectively. In the BE06 corpus no *southward* or *westward* were found; there appeared one *northward* and two tokens of *eastward*. The less frequent occurrence of these direction words in the BE06 corpus possibly indicates that these are maritime vocabulary. We investigated within the CMEC; there appeared 15 times *eastward*, six times *westward* and *northward*, and twice *southward*. So it appears conclusive that these direction-related words can be considered as maritime vocabulary.

It is also worth noticing that the usages in parts of speech differ between the VBC versus the CMEC and the BE06 corpus are pretty different, though, in terms of parts of speech that they are delivered. For instance, 11 tokens of *eastward* in the VBC, as shown in Figure 1 below, are used exclusively as a noun.

1	, was, by a very strong flood tide, drawn to the	**eastward**	of the Island, where she was co
2	extensive sound, with a small arm leading to the	**eastward**.	This was called Point Grey, anc
3	morning, the 10th, having a fine breeze from the	**eastward**,	they stood across Queen Charl
4	is situated about a league to the south	**eastward**	of the presidio of Monterrey. The
5	was dispatched to continue the survey of the coast	**Eastward**	from Cape Hinchinbrook, whilst
6	takes a direction N. 39' W. and the land to the	**eastward**	S. 8' E. It was now the Captain's
7	the ocean, but were obliged to stand to the south	**eastward**,	the wind gradually veering to th
8	clear: there was a considerable swell from the	**eastward**,	and no soundings could be gai
9	sugar loaf hill S. 84 E. and the extreme point to the	**eastward**,	which formed an abrupt Cape,
10	Captain, one afternoon, observed the hills to the	**eastward**	of the river to be on fire, from a
11	, Sunday morning, 18th March, steering to the	**eastward**	or northward, as the wind veere

Figure 1. Concordance of *eastward* in the VBC

While the direction-associated *X-wards* in the VBC are used solely as nouns, *X-wards* in the CMEC and the BE06 corpus show more diverse parts of speech as illustrated in (3) and (4).

(3) *eastward* in the CMEC

a. While there can be no denying the gradual **eastward** shift in shipping's overall centre of gravity, it is (…) [adjective]

b. (…) constituted by the 30°N parallel from Florida **eastward** to 77°30´W meridian, [noun]

c. (…) that wind particles moving toward the equator would come from a region of lower **eastward** velocity and enter a region of higher **eastward** velocity [adjective]

(4) X-ward in the BE06 corpus

a. They set off **northward**. [adverb]

b. In Colchester the population grew, and houses sprawled far beyond the old town walls - into the countryside to the south, and **eastward** as far as the River Colne [adverb]

c. (…) the expansion **eastward** and the remaking of Europe's relation with Washington [noun]

So we conclude that *X-ward* in the 18th and 19th centuries tended to be used as a noun; whereas in contemporary English more diverse parts of speech (noun, adverb, adjective) seem to be presented.

4.3.1.3. Weather-Associated Keywords

This subsection discusses three weather-describing keywords: *pleasant, boisterous,* and *unfavourable*. At first glance, pleasant seemed to describe objects or circumstances; however, concordance for the VBC reveals that eight out of 11 describe the weather (See concordance in Figure 2): the word *weather* as in (5a) and wind in (5b).

1	of the year, very indifferent and exorbitant. With a	**pleasant**	wind, and smooth sea, and fine weather,
2	very variable, and the weather was in general	**pleasant,**	but their progress was considerably
3	with a gentle gale from the N. W. a smooth sea, and	**pleasant**	weather: and now it may be said their
4	Chatham was directed to lead; at this time they had	**pleasant**	weather and a gentle breeze, but it soon
5	weather, were however ill founded; it still continued	**pleasant,**	with a gentle gale chiefly, from the
6	on Tuesday afternoon, the 15th, having had very	**pleasant**	weather, during their excursion, but very
7	21st, the weather was very variable: it then became	**pleasant;**	and they proceeded northward, being on
8	the eastward or northward, as the wind veered, with	**pleasant**	weather, but with such a gentle breeze,
9	mountains (some covered with snow) wearing a	**pleasant**	fertile appearance: along this shore they
10	journey across the continent of America as	**pleasant**	as could be expected from the nature of
11	country was lively, and their journey altogether very	**pleasant.**	On their arrival at the entrance of the

Figure 2. Concordance of *pleasant* in the VBC: all describing weather

(5) a. (...) at this time they had **pleasant weather** and a gentle breeze, but it soon changed, became thick,
b. With a **pleasant wind**, and smooth sea, and fine weather, they lost sight of the Canaries,

As our intuition does not match the result, we performed concordance analysis for the BE06 corpus and the CMEC. A total of 19 and 15 tokens appeared in the BE06 corpus and the CMEC, respectively. None of them are associated with weather or weather-related words. It seems that *pleasant* is an 18th and 19th century domain-specific maritime word to describe the weather[3].

Another interesting weather-describing keyword is *boisterous*. There appeared seven tokens in the VBC, three in the BE06 corpus, and two in the CMEC. All seven tokens in the VBC are used to describe the weather, as illustrated in Figure 3; tokens in the BE06 corpus are not associated with the weather at all but depicting a place as in (6a) or children in (6b). One of the two tokens in the CMEC comes with the weather in (7a), whereas another describes mischievous teenagers in (7b).

1	course to False Bay, imagining Table Ray at this	**boisterous**	season of the year, not only unpleasant
2	the parrallels of 34° 24' and 38° 20' S. lat. but the	**boisterous**	weather prevented him from
3	to the Pacific Ocean. Their apprehensions of	**boisterous**	weather, were however ill founded; it
4	contended with so violent a tempest and such	**boisterous**	weather. As they encreased their
5	their being fatal to unguarded mariners, during the	**boisterous**	weather , which prevails in their vicinity.
6	ill be spared. The weather became exceedingly	**boisterous**,	and the sea broke with so great
7	of Smith's inlet, which they reached on the l4th, in	**boisterous**	rainy weather; and on the 16th, entered

Figure 3. Concordance of *boisterous* in the VBC: all describing the weather

(6) Usage of *boisterous* in the BE06 corpus: no association with the weather
a. **EndPub** itself was never thought of as particularly **boisterous**.
b. (…) if any of **the children** became too **boisterous** but, after a while, even she began to…
(7) Mixed usage of *boisterous* in the CMEC
a. She has experienced rough and **boisterous weather** on the voyage.

3) In order to confirm this postulation, we will have to examine the contemporary English by employing contemporary British English corpus such as British National Corpus, which is beyond the scope of our research purpose. We leave this for future studies.

b. Randall Wells and his team have observed groups of juvenile male bottlenose dolphins behaving like **boisterous teenage** boys.

It seemed that *boisterous* in the target corpus seemed to be exclusively used in the domain-specific maritime word to describe the weather.

4.3.1.4. Anchor-Associated

Five types of anchor-associated keywords appeared, among which we discuss two interesting cases: *bore* and *soundings*. The word *bore*, the past tense form of *bear*, appeared 17 times in the VBC. It is used to locate landscape, in most cases, in conjunction with "by compass", as illustrated in (8). No *bear/bore* was found in the BE06 corpus.

(8) Usage of *bore* in the VBC
a. Mount Baker **bore**, by compass, N. 22 E.
b. The points of the bay **bore**, by compass, S. 32W. and N. 72 W.
c. The land, which in the morning **bore** east and now **bore**, by compass N. 87 E. 8 miles.

The word, *bore* in the CMEC occurred six times, but it is a noun that refers to the diameter of an engine in engineering, as seen in (9).

(9) Usage of *bore* in the CMEC
a. (…) the EEDI limits installed power and so induces owners to use small-**bore** high-rpm engines
b. (…) the coolant (oil) will enter through the piston rod **bore** and will leave through the inside return pipe.

In short, *bear* that has the meaning of "to lie" occurs only in the VBC, and we conclude that *bore/bear* seems to be an 18th and 19th maritime domain word[4].

Another anchor-associated word that occurred in the VBC is *soundings*. The word as

4) As indicated by Assistant Professor Robert J. Dickey, who has been working for Keimyung University and has proofread this paper, *bore/bear* is still used when giving directions to drivers, e.g., "at the 'Y' intersection, bear left," and when identifying incoming attacks of ships or aircraft, e.g., "enemy flight bearing east-northeast at 25,000 feet." We would like to thank Professor Dickey very much for his valuable comments and proofreading.

the meaning of "measurement of depth, especially with a sounding line", according to the definition of Merriam-Webster Dictionary (https://www.merriam-webster.com/dictionary/sounding), occurred six times in both the VBC in (10) and the CMEC in (11).

(10) Usage of *soundings* in the VBC
  a. It was nearly calm during the night, and though within three or four leagues of the land, no **soundings** could be gained at the depth of 130 fathoms.
  b. They tried for **soundings** several times, but could not touch bottom, at the depth of 180 fathoms
(11) Usage of *soundings* in the CMEC
  a. (…) on failure to sight land, or a navigation mark or to obtain **soundings** by the expected time.
  b. (…) if, unexpectedly, land or a navigation mark is sighted or a change in **soundings** occurs.

No tokens in the BE06 corpus were found. *Soundings* seems to be used exclusively in the maritime domain both in the 18th/19th century and contemporary times.

4.3.1.5. Measurements

Two keywords associated with measurements appeared: *fathoms* and *leagues*. As for *fathoms*, a fathom may be equivalent to 6 feet or 1.8288 meters.[5] A total of 41 tokens occurred in the VBC, as some illustrated in (12) and 22 tokens in the CMEC as in (13). Either of the usages is not different and straightforwardly makes reference to the measurement.

(12) a. About 4 o'clock, they had 38 **fathoms**, sand and broken shallow bottom
  b. (…) no bottom could be gained in 180 **fathoms**.
(13) a. Anchor chains are made up of lengths of 15 **fathoms** each.
  b. More shackles must be put in the water: when the water is very deep (more than 25 **fathoms**) in adverse weather…

5) This is not always true because there are at least two different fathoms at sea, and another (different) on land. Thanks for Professor Dickey's comment.

No tokens were found in the BE06 corpus. A *fathom* seems to be a nautical measurement, of which usage is decreasing in contemporary maritime English but is still being used.

A *league* refers to "various units of distance from about 2.4 to 4.6 statute miles (3.9 to 7.4 kilometers)", according to the definition of Merriam-Webster Dictionary (https://www.merriam-webster.com/dictionary/league). A total of 20 tokens appeared in the VBC, as some examples are illustrated in (14).

(14) a. This was called Point Grey, and is about seven **leagues** from Point Robers.
b. After sailing about 10 **leagues**, they came abreast of a small sandy bay.

Only one token in the CMEC and no tokens with the same meaning in BE06 occurred. A *league* as a nautical measurement seems to be obsolete.

4.3.2. Theme-Specific Content Words

As for content words, by employing a bottom-up approach, we divided them into two sub-categories: Proper nouns/titles and theme-related content words. The keywords in this category tell many stories by themselves in terms of "who went where, met whom, and did what". In what follows, we discuss each category.

Let us compare the type number of TD keywords with that of CF keywords in a list of the unshared theme-specific content keywords, as in Table 7 below.

As observed in Table 7, in a category of theme-specific content words, the type number of unshared TD keywords (13 types) outnumbers that of CF keywords (9 types). Furthermore, these TD keywords represent more concrete content-distinctiveness than corresponding CF keywords. That is to say; there are two unshared TD keywords (*commandant* and *gentlemen*) to specify titles of navy position more concretely, whereas there is one general title of the man (*Mr*) occurring as unshared CF keywords, and there are seven theme-related content words (*excursion, admiralty, countrymen, refreshments, habitations, cordiality, provisions*) in a list of unshared TD keywords, whereas there one theme-related content word (*Spanish*) in a list of unshared CF keywords. As compared in Table 7, following a heuristic approach point of view, this cross-comparison also leads us to conclude that text dispersion-based keyword analysis should be much preferred over corpus frequency-based keyword analysis.

Table 7. Comparison of the unshared theme-specific content keywords lists between TD and CF

No	Ranking	TD Keywords	Ranking	CF Keywords
1	23	EXCURSION	34	MR
2	30	ADMIRALTY	39	NAME
3	43	COUNTRYMEN	45	SPANISH
4	50	PUGET	53	HERGEST
5	51	MENZIES	71	COLEMAN
6	52	REFRESHMENTS	81	POMURREY
7	76	CHATHAM'S	82	DALUS
8	79	OWHYHEE	89	OTOO
9	80	COMMANDANT	97	OFFICERS
10	83	HABITATIONS		
11	85	CORDIALITY		
12	91	PROVISIONS		
13	98	GENTLEMEN		

4.3.2.1. Proper Nouns and Titles

A total of 16 type words are classified into this category. The category includes names of people, places, and ships, as well as titles of position. The keywords of each sub-category are presented in Table 8.

Table 8. Keywords regarding proper nouns and titles

Sub-categories		Keywords
Titles	6	captain, commander, commandant, lieutenant, gentlemen, signior
Names of people	5	Broughton, Menzies, Puget, Quadra, Vancouver
Names of Lands of natives	2	Nootka, Owhyhee
Names of ships	3	Discovery, Chatham('s)
	16	

Names of people/ships and titles are intertwined; *Vancouver* is the *Captain* and *commander* of the *Discovery*; *Broughton* is also the *Lieutenant* and *commander* of the *Chatham*; *Signior Quadra* is the *commander* in chief of the Spanish marine; *Puget* is a *lieutenant* of the

Chatham; *Menzies* is a surgeon of the ship and called as a *gentleman*, who in the VBC refers to a social status, unlike contemporary usage. *Nootka* is the land or sound located in the Westcoast of Vancouver Island; *Owhyhee* refers to Hawai'i in the late 18th and early 19th centuries.

We examine whether the above keywords help us to figure out the voyage. See the purpose of the exploration of the Discovery and the Chatham that was presented at the beginning of the text:

> This voyage was undertaken at his Majesty's command, principally with a view to ascertain the existence of any navigable communication between the North Pacific and North Atlantic Oceans, and to make (…) an amicable adjustment of the disputes which had subsisted between the Courts of Madrid and London, relative to the possession of Nootka Sounds, and territory" (Vancouver & Broughton, 1802: 1)

The keywords reveal quite a lot of information regarding the voyage. Specifically, Captain Vancouver, the commander of Discovery and Lieutenant Broughton, along with the second lieutenant Puget and Mr. Menzie, had meetings with Signior Quadra to discuss territorial issues concerning Nootka sounds and territory between Great Britain and Spain. Also, they visited Nootka and Owhyhee a few times. Captain Vancouver visited Hawaii (Sandwich Islands or Owhyhee) three times. And he made two round trips to Owhyhee while he stayed in Nootka to solve an unfortunate event of his crews being murdered in Owhyhee.

4.3.2.2. Theme-Related Content Words

The following 12 theme-related content words reveal more stories regarding the exploration.

(15) theme-related content words
Admiralty, chiefs, cordiality, countrymen, excursion, habitations, Indians, inhabitants, natives, navigators, provisions, refreshments

Such content words help us infer navigators' performance during their expedition. With the keywords, we can postulate that the navigators were affiliated to the Admiralty

of Great Britain; they met Indians, the inhabitants, and their chiefs; some countrymen greeted them with cordiality; in the excursion of lands, the navigators did research on inhabitations and purchased provisions and refreshment. This story is indeed a summary of what the text was written.

In summary, these theme-specific content keywords represent content-generalisability of this target corpus because content-generalisability needs keywords that have the quality to be generalized to the other texts in a similar domain and that "offer insight into the actual content-'aboutness' of those texts" (Egbert & Biber, 2019: 79). As demonstrated in Table 7 and Table 8 above, a requirement of content-generalisability as the other criterion of keywords leads us to conclude that text dispersion-based keyword analysis provides better information than corpus frequency-based keyword analysis.

5. Conclusion

In the present study, a cross-comparison of two keyword analyses has been conducted from a heuristic approach to investigate lexical characteristics of Maritime English used in the 18th and 19th centuries: corpus frequency-based keyword analysis and text dispersion based keyword analysis. Under a cross-comparison of maritime-related CF and TD keywords and theme-specific content CF and TD keywords used in the target corpus, we concluded that text dispersion-based keyword analysis offered better results than frequency-based keyword analysis, in that the former is more effective than the latter regarding the two criteria proposed by Egbert and Biber (2019): the criterion of content-distinctiveness of maritime-related keywords, as well as the criterion of content-generalisability of theme-specific content keywords.

We also conducted a multilateral comparison of maritime-related words between the present target corpus, the VBC, and the BE06 corpus, as well as between the VBC and the CMEC. The findings are summarized as follows:

(i) As a geographical feature, *extremity* is maritime vocabulary that has been being used in its domain.
(ii) As a direction feature, *X-ward* (X = east, west, north, and south) in the 18th and 19th centuries tends to be used as a noun; in contemporary English, more diverse parts of speech (noun, adverb, adjective) seem to be presented.

(iii) As a weather feature, *pleasant* is an 18th and 19th century domain-specific maritime word, not a contemporary maritime word to describe the weather.
(iv) As a weather feature, *boisterous* in the target corpus seems to be exclusively used in the domain-specific maritime word to describe the weather.
(v) As an anchor-associated feature, *bore/bear* that has the meaning of "to lie" seems to be an 18th and 19th maritime domain word.
(vi) As anchor-associated feature, *soundings* seems to be used exclusively in the maritime domain both in the 18th/19th century and contemporary times.
(vii) As a measurement feature, a *fathom* seems to be a nautical measurement, of which usage is decreasing in the CEMC but is still being used.
(viii) As a measurement feature, a *league* as a nautical measurement seems to be obsolete because only one token in the CMEC and no tokens with the same meaning in the BE06 corpus occurred.

This present study faces a limitation based on corpus size. Future steps should include compilation of a much larger sized corpus of navigational expedition journals and logbooks, such that we would be able to confirm our claim that text dispersion based keyword analysis should be much preferred over corpus frequency based keyword analysis. Additionally, an experiment based on a hybrid approach (combination of frequency and text dispersion) to keyword analysis proposed by Gries (2021), comparing the results of text dispersion with those of his hybrid approach, would appear to be indicated in future studies.

References

Allen, D. W. (2002). The British Navy rules: Monitoring and incompatible incentives in the age of fighting sail. *Explorations in Economic History, 39*(2), 204-231.

Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keyword analysis. *Journal of English Linguistics, 3,* 346-359.

Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics, 14*(3), 312-337.

Baron, A., Rayson, P., & Archer, D. (2009) Word frequency and key word statistics in historical corpus linguistics. *International Journal of English Studies, 20*(1), 41-67.

Bertels, A., & Speelmam, D. (2013). 'Keywords Method' versus 'Calcul des Spécificités'. *International Journal of Corpus Linguistics, 18*(4), 536–560.

Bondi, M. (2010). Perspectives on keywords and keyness: An introduction. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 1-18). Amsterdam: John Benjamins Publishing Company.

Chung T., & Nation, P. (2004). Identifying technical vocabulary. *System, 32*, 251-263.

Dunning, T. (1993). Accurate methods for the statistics of surpise an coincidence. *Computational Linguistics, 19*(1), 62-74.

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora, 14*(1), 77-104.

Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and technique. In C. Taylo & A. Marchi (Eds.), *Corpus approaches to Discourse: A critical review* (pp. 225-258). Oxford: Routedge.

Gabrielatos, C., & Sarmento, S. (2006). Central modals in an aviation corpus: Frequency and distribution. *Letras de Hoje, 41*(2), 215-240.

Gelus, J., & Hirch, R. (2019). The reference corpus matters: Comparing the effect of different reference corpora on keyword analysis. *Register Studies, 1*(2), 209-242.

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

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

Gries, S. T. (2021). A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics, 9*(2), 1-33.

Ha, M. (2020). Exploring academic vocabulary in advertising English corpus. *Studies in Linguistics, 57*, 223-2453

Hong, S.-C., & Jhang, S.-E. (2010). The compilation of a Maritime English corpus for ESP learners. *Korean Journal of English Language and Linguistics, 10*(4), 963-985.

Hou, H. (2014). Teaching specialized vocabulary by integrating a corpus- Based approach: Implications for ESP course design at the university level. *English Language Teaching, 7*(5), 26-37.

Jhang, S., Yu, Y., & Lee, S. (2017). Trends in maritime safety standards through keyword analysis of the SOLAS Convention. *Journal of Language Sciences, 24*(2), 227-251.

Lee, S. (2016). *Network analysis of Maritime English Corpus with multi-word compounds: Keyword networks and collocation networks.* Unpublished doctoral dissertation, Korea Maritime and Ocean University.

Liu, S. (2021). *Trends in maritime environment protection through keyness analysis of the MARPOL convention*. Unpublished doctoral dissertation, Korea Maritime and Ocean University.

Qi, Y. (2019). *Syntactic and semantic patterns of domain-specific multiword units in marine accident investigation reports*. Unpublished doctoral dissertation, Korea Maritime and Ocean University.

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

Scott, M. (2016). *WordSmith Tools version 7.* Stroud: Lexical Analysis Software.

Scott, M. (2020). *WordSmith Tools version 8*: Lexical Analysis Software. Retrieved April 15, 2020, from https://www.lexically.net/wordsmith/ downloads/

Scott, M., & Tribble, C. (2006). *Textual patterns, key words and corpus analysis in language education.* Amsterdam, The Netherlands: John Benjamins.

Srichai, P. (2017). A corpus-basd study of specialized vocabulary from American political news articles: An analysis of lexical items. Unpublished master's thesis, Thammasat University.

Stubbs, M. (2010). Three concepts of keywords. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 21-42). Amsterdam: John Benjamins Publishing Company.

Vancouver, G. (Capt), & Broughton, (Lt). (1802). *A narrative or journal of a voyage of discovery to the North Pacific Ocean and round the world performed in the years 1791, 1792, 1793, 1794 and 1795*. London: Lee.

Xu, L. (2021). *Developing vocabulary lists in specialized Maritime English corpora*. Unpublished doctoral dissertation, Korea Maritime and Ocean University.

Xu, L., & Jhang, S.-E. (2020). Keyword analyses of English charter parties. *Linguistic Research, 37*(2), 267-288.

Sunghwa Lee
Associate Professor
Department of English Philology
Korea International University in Ferghana
90 Alisher Navoiy Street, Ferghana City, Uzbekistan
Email: esunghwa@gmail.com

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

Qi Yang
Lecturer
Department of English
School of Foreign Languages
Dalian Maritime University
1 Linghai Rd, Ganjingzi District, Dalian, Liaoning, China
PhD student
Department of English Language and Literature
Korea Maritime and Ocean University
727 Taejong-ro, Yeongdo-Gu, Busan
Republic of Korea, 49112
Email: oxygen888@126.com

Received on August 17, 2021
Revised version received on September 14, 2021
Accepted on September 30, 2021