DOI: 10.53469/jssh.2025.7(02).02

A Corpus-based Study on the Lexical Features of Annual Reports of Shipping Companies: A Case Study of COSCO Shipping and Maersk

Dan Zhang[1,*], Yuling Liang[2], Yating Yang[3]

[1]College of Foreign Languages, Shanghai Maritime University, Pudong, Shanghai, China
[2]Guangxi Huaxi Group Co., Ltd, Nanning, Guangxi, China
[3]Cisco China Co., Ltd. Baoshan, Shanghai, China
[1]danzhang@shmtu.edu.cn, [2]13977816920@163.com, [3]yangyating0403@hotmail.com
**Correspondence Author Dan Zhang*

Abstract: ***As a special type of business text, the annual report has an important function of presenting the development of the company and is an important basis for planning the company's development in the coming year. In this paper, COSCO Shipping and Maersk's annual reports were selected as research objects, and a self-constructed corpus was built to compare and analyse their high-frequency words, lexical density and lexical sophistication. The results found that: (1) The high-frequency words used in the annual reports of COSCO Shipping and Maersk are generally similar, with differences in word frequency and individual words. (2) The lexical density of COSCO Shipping's annual report is lower than that of Maersk's annual report, which discloses a greater amount of information. (3) Maersk's annual report has a higher level of lexical sophistication and is more difficult to read. Based on this, the author also discusses the factors behind the differences, aiming to help English learners and shipping practitioners to better read and understand the annual reports of shipping companies.***

Keywords: Corpus, Lexical features, Annual reports, Shipping companies**.**

1. Introduction

1.1 Research Background

The shipping industry is an important guarantee for international trade and a solid foundation for promoting economic restructuring. Playing an irreplaceable role in promoting the development of the world economy, it provides convenient transport services for China's foreign trade development as a major trading power in the world. There are fluctuations of the industry related to the overall economic, trade, social and policy and regulatory developments at home and abroad. Investors, policy makers and practitioners from the shipping industry read the annual reports to understand the situation of the shipping market and the development of shipping companies. However, in previous research on business text analysis, there were a few researchers have used a corpus approach to study the lexical features of annual reports of shipping companies.

Vocabulary is a core area of knowledge in foreign language studies. For a long time, the study of linguistic vocabulary in the past has been difficult to handle and collect large amounts of linguistic data for technical reasons, so it has had to rely on an introspective approach that relies on intuition as a research method. Nowadays, however, with the advent of the era of big data, the collection of linguistic materials and the processing of textual data have become easier. McEnery & Hardie (2013) argue that corpus linguistics has gradually become one of the mainstream approaches to linguistic research since 1990, and Tuebert (2005) argues that corpora have become the default resource for any language researcher.

Shipping company annual reports, as a special kind of business text, are not only widely used in the industry, but more importantly, the word volume of annual reports meets the basic requirements for building a small corpus of business English. Therefore, using a corpus to study the lexical features of corporate financial annual reports in the shipping industry is certainly a worthwhile option.

1.2 Research Purpose and Significance

In China, maritime transport corridors have become increasingly open and the shipping industry has flourished since the initiative of building a 21st Century Maritime Silk Road was launched in 2013. The booming shipping industry has brought tangible benefits to the development of all countries and has stimulated greater potential for economic and trade cooperation around the world.

For investors, policy makers and practitioners in the shipping industry, the company's financial annual report reflects the operation and finance overview which help them to understand the developments and trends in the industry. For business English learners, they can get better understanding of the shipping industry and broaden their knowledge. It is important for readers to understand and familiarize themselves with the high-frequency keywords of industry annual reports and so as master their lexical characteristics.

Based on the annual reports of two Chinese and foreign shipping enterprises within ten years as a corpus, this paper conducts a comparative analysis on the lexical features of their annual reports. The analysis will explore four perspectives: high-frequency word, lexical density and lexical sophistication, and explore the possible reasons for these differences between the two corpora. Considering that the lexical features of annual reports of shipping companies have not been studied previously, the author is aiming to provide some help to scholars for further research on the lexical features of annual reports of shipping companies in the future.

1.3 Research Questions

This paper uses the corpus analysis software Antconc and Range to study the lexical features of COSCO Shipping and Maersk's annual reports with a comparative approach to explore how the lexical features of the two differ in terms of high frequency words, lexical density and lexical complexity. Based on the findings, we then explore what factors are behind these differences. This study focuses on answering two questions:

(1) What are the lexical features of the annual reports of shipping companies (COSCO Shipping and Maersk as examples)? Are there any differences?

(2) If so, are there any factors behind the differences? In what way do they make the impact?

2. Methods

This chapter will describe the research methodology used in this paper. Firstly, the theoretical basis of the study is presented; then the corpus used in the study is described, and finally the process of data collection, the corpus processing tools used and the process of processing are described.

2.1 Corpus-based Approach

Corpus-based language portrayal is different from introspective or intuition-based language portrayal in that it is objective and reliable. McEnery (2006) outlines the characteristics that a corpus must have:

(1) it must be a machine-readable electronic text;
(2) it must be an authentic occurrence of language;
(3) it must be a strictly sampled language sample;
(4) it is intended to represent a language or language variant. It follows that the corpus must be authentic and representative.

2.2 Corpora Used in the Study

COSCO Shipping and Maersk are the most representative companies in the Chinese and Danish shipping finance sectors based on the top 100 global liner shipping companies in terms of capacity in 2021 (based on the latest capacity data from Alphaliner as of 20 November 2021). The annual reports of Maersk, the top-ranked Danish shipping finance company, and COSCO Shipping, the third-ranked Chinese shipping finance company, for the 10 years from 2011 to 2020 are used as the data source. Invalid data such as data and tables in the annual reports will be removed and English characters will be selected as the data processing part of the corpus.

2.2.1 COSCO Shipping Annual Report Corpus

The selected annual reports were taken from its official website http: //www.coscoshipping.com/. To ensure the completeness of the linguistic material, the corpus covers the English text except for tables, data and images, and contains 623, 466 words, of which, each annual report contains approximately 62, 000 words. The corpus of COSCO Shipping Annual Reports constructed in this study is hereafter abbreviated as CSARC.

2.2.2 Maersk Annual Report Corpus

The selection of Maersk's corporate annual reports as a corpus is representative and has certain research value. This study includes the texts of Maersk's annual reports from 2011 to 2020 from the official website of Maersk (https: //www.maersk.com.cn/), containing 513, 034 English words, of which, each annual report contains about 51, 000 words. The Maersk Annual Reports Corpus constructed in this study is hereafter abbreviated as MARC.

2.3 Data Collection

2.3.1 Tools for Data Collection

Since this paper will analyse the lexical features of the corpus under study from three perspectives: high frequency words, lexical density and lexical complexity, here is a brief introduction to the text processing and corpus data processing software that will be used for this study.

(1) AntConc

AntConc is a powerful green tool, developed by Japanese scholar Laurence Anthony, which can search words, count word frequencies and generate word lists. AntConc makes it easy to count the frequency of words in English texts and to rank the words in order of their frequency of occurrence in the text, and to export the results. In this paper we use Antconc version 3.5.9 to pre-process the text of a self-constructed corpus to produce a morphology-reduced word list.

(2) Range

Range is a software designed by Paul Nation and Averil Coxhead of the School of Linguistics and Applied Linguistics at the University of Victoria, New Zealand. The basic principle is to compare the vocabulary of the text under study with authoritative glossaries to find out how often certain words appear in the glossary, and thus how words are used in the text in general. The BNC is the most authoritative corpus of English, originally created by Oxford University Press in the 1980s and early 1990s, and is an electronic resource of 100 million words on a wide sample of written and spoken language from a wide range of sources, presenting British English from the late 20th century onwards. It covers both spoken and written English. The written corpus is 90% and the spoken corpus is 10%. The corpus is both written and spoken, with a word count of over 100 million words, and consists of 4124 texts representing a wide range of modern British English. The General Service List and the Academic Word List, a list of words compiled by a group of the world's leading experts in English language studies, are also used in this study. The GSL is a list of the most important words for second language learners of English in Europe and the United States, and includes all of the 2, 284 most frequently used words in the world. The AWL was developed by Averil Coxhead, a distinguished professor of linguistics, through corpus research, tracking all mainstream media and academic articles in the English-speaking world over many years, to produce a list of the 570 most frequent English words in all academic articles.

The high-frequency words studied in this paper make use of the Range software's high-frequency word family hierarchical distribution function (based on the BNC, GSL and AWL corpora). Word density, TTR will also be used with Range software to assist in analysing the data results. This section is described in detail in later chapters.

(3) Corpus Word Parser.

Corpus Word Parser is a specialised corpus word division and lexical annotation tool. It is a search-based collocation extraction tool which yields MI, MI3, T-score, Z-score, Log-Log, and Log likelihood scores of collocational strength. The tool works will raw and CLAWS-tagged PoS English texts, and does not work for texts of Chinese or other languages.

2.3.2 Procedures of Data Collection

The collection of data is a prerequisite for analysis. The first step was to build two corpora. The author searched and downloaded the annual reports of both COSCO Shipping and Maersk from 2011-2020, including numerical tables and other figures, for a total of approximately 1, 200, 000 words.

Then the author compiled the language materials. The author converted the downloaded annual reports into word format and eliminated invalid data such as images and tables from each document, and carried out preliminary text cleaning. After carrying out the initial text processing the author summarised the annual report data of the two companies over a ten-year period into two separate word documents.

The two corpora were summarised respectively using the "find and replace" function that comes with office software, in order to "find the content" column, enter "[0-127] ""[, . ;: %]" to eliminate invalid data such as numbers and punctuation from the text in bulk.

Finally, the cleaned corpus data is converted to lowercase format and exported to produce a txt file.

2.4 Procedures of Data Processing

After simple text pre-processing on the targeted data in the defined range, the corpus of Maersk's annual report had 489478 characters and the corpus of COSCO Shipping's annual report had 545355 characters. Corpus Word Parser software for acrostic processing. Finally, the text was imported into Antconc version 3.5.9, and the Lamme List downloaded from the Antconc website was imported into the Tool Preferences section, and the Stop List was imported to produce valid word lists that had been clustered and the common invalid words removed. The generated word lists are arranged in order of frequency.

Table 1: Types and Tokens Statistics for the Two Corpora

Corpus	Type	Token
CSARC	8041	545355
MARC	15095	489478

As Liu's (2017) study concluded that the two-eight law method is more applicable to intercepting domain high-frequency words than the Price's formula selection method, this study will use the two-eight law in the calculation of LS values to define high-frequency word thresholds for the self-constructed corpus. The top 20% of each of the two generated word lists were selected as high-frequency words according to the two-eight law. Subsequent data analysis was based on the attached high-frequency word lists and the actual data analysis needs.

The Type Token Ratio is a valid measure of text complexity, and is a more commonly used measure of lexical density in the academic community. One definition refers to the different words that appear in a text, with different inflections of a word such as learn and study considered as different words. The other, more accurate, definition is to consider the different inflected forms of a word as the same token, e.g. learn and study as the same token. The second definition is used in this study, and the text is pre-processed using the word form reduction software Lemmatizer to achieve more accurate results.

The size of the TTR is influenced by the size of the corpus, with the larger the text, the smaller the value. In this study, the RTTR formula proposed by Guiraud (1960) was used to measure the lexical density of shipping company annual report texts.

Lexical sophstication refers to the ratio of infrequently used words to high level words in the text. Read (2000) proposed a formula for calculating LS, i.e. LS = number of word families of complex words (low frequency words) / total number of word families in the text. However, this calculation method is sensitive to the length of the text, i.e. the longer the text, the lower the complexity. In this paper, the author adopt the findings of Mao-Cheng Liang (2011) and use the ratio of low-frequency words to high-frequency words to measure lexical complexity, using the calculation formula LS = low-frequency words/high-frequency words. For the convenience of data extraction, low-frequency words are defined as real words with a frequency of 1 in the text. In the text, low-frequency words are represented in both their singular and plural forms.

According to the formula:

$$LS = \frac{L\text{-}Freq\ Words}{H\text{-}Freq\ Words}$$

The results were as follows.

Table 4: Lexical Sophstication for the Two Corpora

Corpus	H-Freq Words	L-Freq Words	LS
CSARC	1632	2741	1.68
MARC	3019	7291	2.42

Note: Data in table retained to two decimal places.

3. Results and Analysis

Based on the results of the data analysis in Chapter 3, Sub-section 4, this chapter will first analyse the lexical characteristics of the annual reports of COSCO Shipping and Maersk from three perspectives: high frequency word distribution, lexical density and lexical sophistication, and explore the differences between the two. Finally, the author will explore what factors may have an impact on the differences in lexical features between the two.

3.1 Analysis of the Lexical Features Based on the Data Processing Result

3.1.1 Distribution of High-frequency Words

High-frequency words refer to words that appear more frequently in corpus texts, which can reveal the focus of a text or demonstrate certain features of a chapter.\After importing the two corpora, CSARC and MARC, into Antconc software, two high-frequency word lists were generated using its Word List function. In order to better understand the lexical features of COSCO Shipping and Maersk's annual reports, the CLAWS7 website (http://ucrel-api.lancaster.ac.uk/claws/free.html) was used to lexically assign the generated high-frequency word lists and counted the number and percentage of these lexical features. BNC1994, BNC2014 and all English corpora in Mark Davies' BYU corpus server for POS tagging. This study uses the C5 tag set for lexical tagging. The table below shows the number and percentage of lexical categories for the top 100 high-frequency words in CSARC and MARC.

Table 5: Top100 lexical classification of high frequency words (CSARC)

Word Class	Examples	Total
Noun	company, group, ship, asset, container, vessel, etc.	61
Adjective	financial, general, current, total, etc.	10
Adverb		0
Verb	report, loss, share, control, etc.	9
Others	the, a, for, of, as with, that, etc.	20

Table 6: Top100 lexical classification of high frequency words (MARC)

Word Class	Examples	Total
Noun	income, oil, activity, asset, price, capital, etc.	59
Adjective	financial, general, current, total, etc.	7
Adverb	not	1
Verb	share, supply, hedge, change, etc.	12
Others	the, a, for, of, as with, that, etc.	21

As per the data in the table above, the top 100 high-frequency words in both the CSANC and MARC corpora have the largest proportion of nouns, accounting for 61% and 58%, respectively. The proportion of adjectives and adverbs is 9% and 8% respectively. This indicates that the proportion of nouns containing actual meaning is the largest in the top 100 high-frequency words in both the CSANC and MARC corpora, while adjectives and adverbs with modifying meanings are used less frequently. It can be inferred that nouns are used much more frequently than other lexical words in the texts of COSCO Shipping and Maersk's corporate annual reports. The high use of nouns and the small proportion of modifiers with subjective meanings indicate the official style of COSCO Shipping and Maersk's annual reports. Based on the word list generation function of Antconc software, the author divided the top 100 high-frequency words of both CSARC and MARC corpora into three categories according to semantics and usable contexts, including business-related words, shipping-related words and others.

The tables are as follows.

Table 7: Top 100 High Frequency Words Classification (CSARC)

Groups	Examples	Total
Business-related words	profit, asset, trade, subsidiary, liability, consolidate, ect	44
Ship-related words	ship, container, board, line, vessel, ect	16
Others	the, a, for, with, recognised, december, ect	40
Total		100

Table 8: Top 100 H-Freq Words Classification (MARC)

Groups	Examples	Total
Business-related words	profit, tax, hedgel, liability, revenue, consolidate, ect.	41
Ship-related words	oil, container, terminal, line, vessel, ect	13
Others	the, a, for, with, recognised, december, ect	46
Total		100

From Table 7 and Table 8, it can be visualized that among the top 100 high-frequency words in COSCO Shipping's annual report, both business-related words and shipping-related words are higher than Maersk's, exceeding Maersk's by 6 words year-on-year, accounting for 6% of the total number of words in the table. Based on these data, the author can conclude that the top 100 high-frequency words in the annual reports of COSCO Shipping and Maersk include a large proportion of business-related words. From this, it can be concluded that the vocabulary of COSCO Shipping and Maersk's corporate annual reports has the characteristics of BE. Therefore, readers or practitioners in the shipping industry will need to improve their reading ability of the annual reports of shipping companies by learning to master more BE vocabulary. For the writers of shipping annual reports, it is beneficial to expand their business-related vocabulary to enhance the professionalism of writing shipping annual reports.

By comparing the top 100 high-frequency words in CSARC and MARC, it is found that the high-frequency words are similar in general, with some differences in frequency and diversity. For example, the words oil, net, global and hedge appear in MARC's top 100 list of high frequency words, but not in CSARC's top 100 list. In contrast, the words subsidiary and committee only appear in the CSARC top 100 list. The reason for this difference is related to the business scope and organisational structure of COSCO Shipping and Maersk. COSCO Shipping's main business is the handling and storage of containers and break-bulk cargo terminals. It is a group with a large number of secondary and tertiary subsidiaries. Therefore, it is reflected in the high frequency word list as a subsidiary, with high frequency of committee. And the word container has a higher word frequency than the word 1197 times in the MARC high frequency word list. Maersk's scope of business prior to 2016 was highly energy-related, in which most of its customers' bookings were online. Thus reflecting the high frequency use of the word oil, net. In summary, it can be concluded that the high frequency words in the annual reports of shipping companies reflect the scope of business of the companies. The reader can quickly get an idea of the business scope of the company based on the list of high frequency words in the annual reports of shipping companies.

3.1.2 Lexical Density

According to Halliday (1985), the words in a sentence are divided into grammatical items (words that play a qualifying role in the sentence, such as crowns, pronouns, most prepositions, conjunctions and finite verbs) and lexical items (real words). The lexical density reflects the proportion of text occupied by real words., and a higher lexical density means more real words and more information is covered, and the vice versa.

Harley & King (1989) found that native speakers write texts with greater lexical density than second language learners. Danish is the official language of the Kingdom of Denmark and belongs to the North Germanic branch of the Indo-European-Germanic family of languages, with which it is interchangeable with Norwegian and Swedish. Although neither Denmark nor China is a native English-speaking country, English and Danish belong to the same Germanic group so they share many similar words. Based on Harley & King's findings and the above reasons, it can be conjectured that the vocabulary density of Maersk's annual report is greater than that of COSCO Shipping's annual report.

To test this conjecture, the author used Antconc software to derive the number of types and tokens for both CSARC and MARC corpora, and derived the value of RTTR according to the RTTR formula proposed by Guiraud (1960).

According to Figure 1, the formula for calculating RTTR was brought in:

$$RTTR = \frac{Ttpes}{\sqrt{Tokens}}$$

The results obtained were as follows.

Table 9: Root Type Token Ratio for the Two Corpora

Corpus	Type	Token	RTTR
CSARC	8041	545355	10.89
MARC	15095	489478	21.58

Note: Data in table retained to two decimal places.

As can be seen from the data in Table 3.4, the total number of tokens in COSCO Shipping's corporate annual report is greater than the total number of tokens in Maersk's annual report, but the total number of types is less than the total number of types in Maersk's annual report. COSCO Shipping's RTTR is also lower than Maersk's RTTR. This indicates that the vocabulary density of COSCO Shipping's corporate annual report is lower than that of Maersk's corporate annual report. In other words, Maersk's annual report is more dense and contains more information than COSCO Shipping's annual report. When reading the annual reports of both companies, Maersk's annual report is more difficult to read than COSCO Shipping's.

3.1.3 Lexical Sophistication

The results in Table 3-4 show that Maersk's annual report has a higher number of high-frequency words and low-frequency words than COSCO Shipping's annual report. The higher LS value reflects the fact that Maersk's annual report contains more low-frequency words and is more difficult to read.

The author imported the text data of both annual reports into the Range software based on BNC for Baseword List for analysis, and the results were obtained in the following table.

Table 10: Lexical Distribution of CSARC in BNC

Word List	Type/%	Token/%	Families
one	372802/67.67	1950/19.65	763
two	67018/12.17	1157/11.66	530
three	12418/ 2.25	433/ 4.36	255
four	24830/ 4.51	513/ 5.17	270
five	9470/ 1.72	310/ 3.12	187
six	3927/ 0.71	170/ 1.71	120
seven	3595/ 0.65	112/ 1.13	86
eight	1114/ 0.20	79/ 0.80	72
nine	516/ 0.09	59/ 0.59	44
ten	486/ 0.09	35/ 0.35	30
11	2362/ 0.43	28/ 0.28	26
12	319/ 0.06	27/ 0.27	22
13	155/ 0.03	21/ 0.21	19
14	402/ 0.07	25/ 0.25	22
15	6332/ 1.15	137/ 1.38	137
not in the lists	45146/ 8.20	4866/49.04	????
total	550892	9922	2583

Table 11: Lexical Distribution of MARC in BNC

Word List	Type/%	Token/%	Families
one	301812/62.05	2352/11.52	855
two	54322/11.17	1461/7.15	628
three	14105/2.90	656/3.21	357
four	18168/3.74	619/3.03	340
five	9335/1.92	396/1.94	227
six	3504/0.72	245/1.20	171
seven	3615/0.74	161/0.79	124
eight	1323/0.27	127/0.62	106
nine	824/0.17	91/0.45	67
ten	634/0.13	90/0.44	72
11	1125/0.23	45/0.22	39
12	402/0.08	52/0.25	44
13	311/0.06	46/0.23	36
14	127/0.03	27/0.13	27
15	5316/1.09	313/1.53	313
not in the lists	71457/14.69	13739/67.28	????
total	486380	20420	3406

Note: Data in table retained to two decimal places.

According to Table 10 and Table 11, the annual reports of COSCO Shipping and Maersk are mostly distributed in the first and second vocabulary levels of BNC. The number and proportion of real words in the first and second vocabulary levels of COSCO Shipping's annual reports are higher than those of Maersk.

The text data of both annual reports were imported into the Range software based on GSL and WAL for Baseword List for analysis, and the results are shown in the following table.

Table 12: Lexical Distribution of CSARC in GSL and AWL

Word List	Type/%	Token/%	Families
one	372316/67.58	1776/17.90	762
two	22194/4.03	567/5.71	302
three	54485/9.89	1195/12.04	453
not in the lists	101897/18.50	6384/64.34	???
total	550892	9922	1517

Table 13: Lexical Distribution of MARC in GSL and AWL

Word List	Type/%	Token/%	Families
one	298160/61.30	2095/10.26	833
two	20335/4.18	776/3.80	405
three	50270/10.34	1368/6.70	494
not in the lists	117615/24.18	16181/79.24	????
total	486380	20420	1732

Note: Data in table retained to two decimal places.

Based on the data in Table 12 and Table 13, it tells that the total number of real words (tokens) in COSCO Shipping's corporate annual report accounts for a higher percentage of Baseword one and Baseword two (i.e. the General Glossary of English Words GSL developed by West) than Maersk's, with COSCO accounting for 71.61% and Maersk accounting for 65.48%. And the number of tokens in Maersk's annual report is slightly higher than COSCO Shipping's share of Baseword three (i.e. Academic English Vocabulary List AWL), which is 0.45% higher than COSCO's share. These data suggest that COSCO Shipping's corporate annual report vocabulary overlaps more with GSL than Maersk's corporate annual report vocabulary. In contrast, reading Maersk's annual report requires more AWL vocabulary than reading COSCO's annual report. In other words, if the reader knows most of the vocabulary in the BNC or GSL, it is less difficult to read COSCO Shipping's annual report than Maersk's annual report.

In summary, the vocabulary complexity of Maersk's annual report is higher than that of COSCO Shipping's annual report, and it is also more difficult to read than COSCO Shipping's annual report.

3.2 Reasons for Differences between the Two Corpora

Based on the above analysis, the author explores the reasons affecting the differences in the lexical features of COSCO Shipping and Maersk's annual reports. The following three reasons that may affect these differences are also proposed.

3.2.1 Cross-cultural Factors

(1) Power distance

According to the high-frequency word lists of COSCO Shipping and Maersk, the nouns management and manager appear 3238 times in COSCO Shipping's annual report, with the word frequency of management being 1898 and manager being 837, while in Maersk's annual report, management appears 1070 times and manager only 11. COSCO Shipping is under the influence of the Chinese culture with a large power distance, which emphasizes the corporate operation model of top management and subordinate staff execution. Therefore, the vocabulary of the annual report reflects the high frequency of words such as management, manager and executive, which emphasize the position of authority.

(2) Collectivism

According to the high frequency word list generated by Antconc, COSCO Shipping uses the terms "China, Chinese" or "company, group" in a collective sense. Maersk, on the other hand, uses words such as "business, customers". From a cultural point of view, China is a collectivist culture that values the realisation of group interests. At the same time, as a state-owned enterprise, COSCO Shipping has frequent use of words such as "China, Chinese", which is a sign of the company's emphasis on expressing the nature of its national affiliation. Under the influence of the culture of collectivism, terms such as associate and consolidate appear much more frequently in COSCO Shipping's annual reports than in Maersk's. Maersk's use of terms such as 'customers' helps the company to demonstrate its consumer-focused corporate philosophy and to bring it closer to its customers.

(3) Thinking patterns

Because in the West there is more of a linear mindset, whereas in Chinese culture there is a more curvilinear mindset, westerners pay more attention to detail. Western thinking is straightforward and abstract, like a straight cut, with a clear division and an emphasis on abstract reasoning. In contrast, Chinese thinking is like a circle with an internal seal, an overview, a search for an epiphany, a curvilinear thinking, characterised by wholeness and intuition. As a result of this difference in mindset, Maersk's annual report is more specific and contains more information than COSCO Shipping's. Meanwhile, the emphasis on detail has also influenced the complexity of the vocabulary used in Maersk's annual report to a certain extent, making it more complex than COSCO Shipping's.

3.2.2 Language factors

COSCO Shipping Company is a mega state-owned enterprise in China directly managed by the central government, in a society where Chinese is the mother tongue. Maersk, on the other hand, is established in Denmark, a country where Danish is the first language. For historical reasons, Danish has affinities with Norwegian, Swedish and Icelandic, and is particularly similar to Norwegian and Swedish. The official language of Denmark is Danish, which belongs to the East Scandinavian branch of the North Germanic language family. English, on the other hand, belongs to the same Germanic language group, so there are many similar words in both languages. The Danish words have, for example, over, under, for have something in common with their English counterparts, as they are identical or similar in structure to their English counterparts. English is spoken by a large number of people in Denmark and is the second language in the country. Therefore, it can be assumed that the native language used has an influence to some extent on the lexical density and lexical complexity of COSCO Shipping and Maersk's annual reports.

3.2.3 Corporate Business Strategy Factors

The growth strategy and focus of the companies also influenced the differences in the lexical features of COSCO Shipping and Maersk's annual reports. Maersk was an integrated energy shipping company (including Maersk Line and Maersk Tanker) before its strategic shift to a complete "end-to-end" integrated logistics business provider in 2016. In the Top 100 list of high frequency words from Maersk's annual report, oil was mentioned 1, 385 times, while it does not appear in the top 100 high frequency words in COSCO Shipping's annual report. At the same time, the words "change, new, risk, hedge" appear frequently in Maersk's annual report because of the increase in offshore projects due to the plunge in oil prices in 2016. Among the top 100 high frequency words in COSCO Shipping's annual report, there are more words that indicate time, such as "year, December, period, current", as well as many relevant words that emphasise the company's development plans, such as "statement, development, continue". This might be led by

COSCO Shipping's reorganisation of COSCO Group and China Shipping Group in 2015. After the merger of COSCO Group and China Shipping Group to form COSCO Shipping, COSCO Shipping has significantly increased its capacity, moving forward in the global ranking, developed and expanded its business scale, relying on continuous acquisitions and construction of new vessels. The company's growth strategy is aggressive and forward-looking. Under the influence of such corporate development scale and development strategy, COSCO Shipping's corporate annual report high frequency words also show high correlation with time and development plan.

4. Conclusion

4.1 Major Findings

From the above analysis, the vocabulary of COSCO Shipping and Maersk's annual reports has the following characteristics: (1) The high-frequency vocabulary of COSCO Shipping and Maersk's annual reports reflects their business scope, and the frequency of using nouns is higher than the frequency of using other lexical words. Both annual reports have an official style and the vocabulary has the characteristics of BE. (2) The vocabulary density of COSCO Shipping's annual report is lower than that of Maersk's annual report. Maersk's annual report reflects more intensive information and contains more information than COSCO Shipping's annual report. (3) The lexical complexity of Maersk's corporate annual report is higher than that of COSCO Shipping's corporate annual report, which makes it more difficult to read than COSCO Shipping's annual report.

Factors affecting the differences in the vocabulary characteristics of COSCO Shipping's and Maersk's corporate annual reports include cross-cultural factors, language factors and corporate business strategy factors. (1) Under the influence of high power distance and collectivist cultural factors, the high-frequency words in COSCO Shipping's annual report show the characteristics of a state-oriented and authoritative subject, while the vocabulary of Maersk's annual report shows its consumer-oriented attributes. (2) The differences in thinking patterns and language families make the lexical density and complexity of Maersk's annual reports higher, and the information density and reading difficulty higher than those of COSCO Shipping's annual reports. (3) Objective factors in the development of the company and its development strategy affect the frequency of individual words in the annual reports of both companies.

4.2 Implications

For practitioners in the shipping industry and learners of English, the annual reports of listed companies provide a clear picture of the company's operations and developments. Corporate annual reports provide an at-a-glance view of developments and trends in the industry and are very important official business texts. However, reading shipping company annual reports can be difficult and challenging for some English learners who do not have a deep understanding of the industry. For those working in the shipping industry, corporate financial annual reports are in reality a special business text that needs to be consulted frequently. This study analyses the vocabulary of COSCO Shipping and Maersk from three perspectives: high-frequency words, vocabulary density and vocabulary complexity, to help English learners gain a preliminary understanding of the lexical features of the annual reports of shipping companies. At the same time, the study investigates the influencing factors behind the differences in the lexical features of the two companies, which will be of some significance for shipping practitioners to better read, understand and write the annual reports of shipping companies. The author hopes that the findings of this study will provide some implications for future researchers.

4.3 Limitations and Suggestions

Considering that the corpus in this study only covers the annual reports of COSCO Shipping and Maersk within a decade, the sample is limited to these two companies in China and Denmark, and the sample size has room to be expanded. Due to the large volume of data in the corpus, there are inevitable data errors caused by processing omissions when cleaning the text data.

Based on the findings of this study, the author suggests that shipping practitioners and English learners can improve their reading capability of shipping enterprise annual reports by learning more BE vocabulary, and can also increase their understanding of shipping enterprise annual reports by reading more formal shipping business texts. At the same time, English language learners in the shipping industry need to improve their AWL vocabulary in order to better understand and write their annual reports.

The author also suggests shipping practitioners take into account the cultural environment of their company when writing their annual reports. For example, if you are in a culture that values collegiality, you should present a holistic view. If the company is in a culture that prefers a linear mindset, the report should be written in a straightforward manner to disclose specific information. In addition, the vocabulary of a shipping company's annual report reflects the company's main business and development strategy to a certain extent, so English learners can improve their mastery of shipping company annual report vocabulary by understanding the high frequency words used by shipping companies.

Acknowledgement

This research was supported by the "International Maritime English" Specialty Course Project (2024) from Shanghai Maritime University.

References

[1] Aarts, J&Theo van den Heuvel. (1982). Grammars and intuitions in corpus linguistics. In S. Jonhansson (ed.). Computer Corpora in English Language. Norway Computing Center for the Humanities.66-84.

[2] Birgit Harley& Mary Lou King. (1989). Verb Lexis in the Written Compositions of Young L2 Learners. Studies in Second Language Acquisition. 11(4): 415-439

[3] Guiraud, P. (1960). Problemes Et Methodes De La Statistique Linguistique. Dordrecht: D. Reidel.
[4] Halliday. (1985). Spoken and Written Language. Oxford University Press.
[5] Kenneth Hyltenstam. (2010). Lexical characteristics of near-native second-language learners of Swedish. Journal of Multilingual and Multicultural Development, 9(1-2): 1-84.
[6] Lanfer, B&P, Nation. (1995). Vocabulary size and use: lexical richness in L2written production. Applied Linguistics 16: 307-322.
[7] M. A. K. Halliday & Ruqaiya Hasan (1985) Language, Context, and Text: Aspects of language in a social-semiotic pespective. Waurn Ponds, Vic: Deakin University.
[8] McEnery, Tony and Andrew Hardie. (2013). Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press.
[9] McEnery, A. et al. (2006). Corpus-Based Language Studies: An Advanced Resource Book. London & New York: Routledge.
[10] Nation. I. S. P. (2001). Learning Vocabulary in Another Language. Cambridge: Cambridge University Press.
[11] Qian Yufang & Tony McEnery. (2017). A corpus-based discourse study of Chinese medicine in UK national newspapers. Foreign Language Teaching and Research. (01), 73-84.
[12] Sinclair, John. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.
[13] Sonca Vo. (2019). Use of lexical features in non-native academic writing. Journal of Second Language Writing, 2019(44): 1-12.
[14] Tuebert, W. (2005). My version of corpus linguistics. International Journal of Corpus Linguistics, 10(1): 1-13.
[15] Tognini-Bonelli, Ellena (2001). Corpus Linguistics at Work. Amsterdam & Philadelphia: John Benjamins.
[16] Ure (1971). Lexical density and register differentiation. In G. Perren and J.L.M. Trim (eds). Applications of Linguistics. Cambridge University Press.
[17] Zhang Baicheng. (2009). FL Vocabulary Learning of Undergraduate English Majors in Western China: Perspective, Strategy Use and Vocabulary Size. English Language Teaching, 2(3): 178.
[18] Ba Zhichao, Li Gang, Zhu Shiwei. Research on Keyword Selection and Semantic Measurement Methods in Co-occurrence Analysis. Journal of Intelligence, 2016(02), 197-207.
[19] Bu Han. A comparative study on the impact of discourse quality of annual reports of listed companies in China and the U.S. on capital market reactions. Doctoral dissertation, 2019, University of International Business and Economics.
[20] Chen Rong. Cognitive Contextual Research on English Vocabulary Teaching. Doctoral dissertation, 2011, Southwest University.
[21] Liang Maocheng. The construction of automatic scoring model for Chinese students' English compositions. Shanghai: Shanghai Foreign Language Education Press, 2011.
[22] Liu Yishan, Wang Yulin, Li Mingxin. Empirical analysis of the applicability of high-frequency word threshold definition method in word frequency analysis. Digital Library Forum, 2017(09), 42-49.
[23] Wang Lifei, Han Fang. Comparative analysis of language wheels in Chinese and English corporate annual report genres. Journal of PLA College of Foreign Languages, 2015(05), 1-9+107+159.
[24] Wu Nan, Zhang Jingyuan. Discourse Construction Strategies of Corporate Institutional Identity in China and the United States. Modern Foreign Languages, 2019(02), 220-230.
[25] Sun Qinglan. Boundary division between high-frequency words and low-frequency words and word frequency estimation method. Chinese Journal of Librarianship, 1992(02), 78-81+95-96.
[26] Xu Jun, Xiao Haiyan. A study of business translation based on critical genre analysis (CGA). Chinese Foreign Language, 2016(04), 20-28.
[27] Zhang Yanfei. Lexical Characteristics and Translation Strategies of Scientific and Technical Texts. Master's thesis, 2020, Shanghai Jiao Tong University.
[28] Zhu Huimin, Wang Junju. Developmental features of lexical richness in English writing - a longitudinal study based on a self-constructed corpus. Foreign Language Community, 2013(06): 77-86.

APPENDIX

List of Abbreivations

BE: Business English
BNC: British National Corpus
GSL: General Service List
AWL: Academic Word List
TTR: Type Token Ratio
RTTR: Root Type Token Ratio
LS: Lexical Sophistication
CSARC: COSCO Shipping Annual Report Corpus
MARC: Maersk Annual Report Corpus

Table 2: High -frequency Word List Top100 (CSARC)

Rank	Freq	Lemma
1	45774	the
2	29321	of
3	19403	and
4	13023	be
5	12142	to
6	10165	in
7	8392	a
8	7955	company
9	5250	RMB
10	5178	for
11	5069	as
12	4935	group
13	4509	or
14	4404	financial
15	4274	ship
16	3944	on
17	3741	at
18	3383	asset
19	3076	director
20	3033	year
21	3026	with
22	2947	from
23	2891	december
24	2835	have
25	2679	china
26	2475	by
27	2471	container
28	2400	that
29	2394	other

30	2299	lease
31	2142	report
32	2125	Mr
33	2123	loss
34	2047	statement
35	2038	share
36	2008	interest
37	1945	note
38	1912	value
39	1898	management
40	1841	continue
41	1824	Ltd(Limited)
42	1822	its
43	1787	board
44	1769	co
45	1762	cash
46	1760	finance
47	1721	which
48	1690	amount
49	1671	profit
50	1664	liability
51	1613	limit
52	1598	service
53	1562	investment
54	1498	general
55	1455	transaction
56	1445	consolidate
57	1443	executive
58	1400	fair
59	1374	meeting
60	1340	committee
61	1340	control
62	1301	cost
63	1278	business
64	1273	tax
65	1266	end
66	1249	income
67	1223	net
68	1221	under
69	1214	rate
70	1183	risk
71	1178	annual
72	1155	not
73	1153	any
74	1149	subsidiary
75	1111	line
76	1079	include
77	1076	use
78	1051	period
79	1040	account
80	1037	vessel
81	1033	COSCO
82	1021	equity
83	1018	party
84	1016	term
85	958	development
86	944	current
87	928	recognised
88	909	operation
89	907	trade
90	907	will
91	885	associate
92	870	operate
93	864	price
94	855	total
95	847	relate
96	839	impairment
97	839	no
98	837	manager
99	823	independent
100	821	corporate

Table 3: High-frequency Word List Top100 (MARC)

Rank	Freq	Lemma
1	28079	the
2	18334	of
3	16839	and
4	16098	be
5	13183	in
6	11379	to
7	8667	a
8	5737	for
9	4741	usd
10	4510	on
11	4242	Maersk
12	3896	as
13	3376	by
14	3352	share
15	3331	financial
16	2941	with
17	2655	from
18	2495	have
19	2331	company
20	2295	other
21	2288	billion
22	2280	at
23	2253	asset
24	2194	value
25	2137	group
26	2038	ap
27	1938	year
28	1907	statement
29	1871	DKK(Danish Krone)
30	1865	cost
31	1820	rate
32	1811	cash
33	1774	board
34	1771	report
35	1762	total
36	1751	use
37	1740	loss
38	1644	tax
39	1608	terminal
40	1516	risk
41	1416	include
42	1414	activity
43	1408	that
44	1385	oil
45	1370	amount
46	1366	increase
47	1365	income
48	1336	or
49	1333	business
50	1319	flow
51	1293	service
52	1283	note
53	1279	which
54	1274	container
55	1240	profit
56	1217	interest
57	1204	capital
58	1173	net
59	1134	annual
60	1126	relate
61	1097	fair
62	1081	sale
63	1079	price
64	1070	management
65	1057	not
66	1054	plan
67	1047	this
68	1011	will
69	1010	etc
70	1009	liability
71	994	operate
72	986	hedge
73	986	recognised
74	981	our
75	954	market
76	913	base

77	913	revenue
78	903	change
79	899	current
80	884	vessel
81	849	continue
82	834	customer
83	803	consolidate
84	802	december
85	802	per
86	800	trade
87	799	operation
88	795	director
89	786	supply
90	782	million
91	772	line
92	766	end
93	754	new
94	739	time
95	737	lease
96	735	global
97	727	currency
98	721	impairment
99	706	impact
100	704	gain