A Study on the Lexical Complexity of English for Special Purpose Based on WordSmith Tools and Software Range

A Case Study of Medical English, Maritime English and English for Science and Technology

Minghui Nie

School of Foreign Studies, Nanjing University of Science and Technology, Nanjing, Jiangsu, China, 210094

nmhnjust@njust.edu.cn

Abstract. This study compared the lexical complexity of different English for Special Purpose (ESP), mainly based on the Marine Engineering English Corpus (MEE), Medical English Corpus (MEC) and Jiao Da English for Science and Technology Corpus (JDEST). The study made a general analysis of word length, lexical sophistication of lexical frequency files and hapax legomena, and lexical diversity of the ratio of types and tokens generated by Wordsmith Tools and Range. The research results are as follows: (1) Medical English has significantly longer words than Maritime English, while Maritime English surpasses English for Science and Technology; (2) regarding vocabulary sophistication, Medical English and English for Science and Technology demonstrates higher sophistication than Medical English; (3) English for Science and Technology exhibits significantly higher lexical diversity compared to Medical English, which in turn is slightly higher than Maritime English.

Keywords: Lexical complexity, English for Special Purpose, WordSmith Tools, Software Range.

1 INTRODUCTION

1.1 English for Specific Purposes and Lexical Complexity

With the rapid development of the global economy and technology, English teaching is shifting from general to professional English. English for Specific Purposes (ESP) includes English for academic, professional, and occupational purposes. ESP teaching philosophy and methods differ from general English, aiming to equip learners with English skills in specific fields for practical communication.

This study examines medical English, maritime English, and scientific English to compare their characteristics and provide learning and teaching suggestions. Zhu [9] summarized medical English in terms of vocabulary, sentence structure, and grammar. Lyu and Gu [2] analyzed Maritime English conjunctions focusing on lexical

M. S. H. Talpur et al. (eds.), *Proceedings of the 2024 4th International Conference on Internet Technology and Educational Informatization (ITEI 2024)*, Atlantis Highlights in Social Sciences, Education and Humanities 27,
https://doi.org/10.2991/978-94-6463-560-7_10

density, word frequency, location distribution, and semantic distribution using a corpus. Huang and Yang [1] used a computer and the JDEST corpus to analyze English used by mechanical and electronic majors, covering word length, frequency, and text distribution.

Studies on lexical complexity often relate to second language teaching. Wolfe-Quintero et al. [6] highlighted the importance of having a variety of basic and sophisticated words. Wei [5] uses corpora to analyze data is common, as it facilitates retrieval and analysis, benefiting classroom teaching and research. Zhao [7] compared vocabulary frequency files to understand lexical complexity in different corpora. Vermeer [4] suggested that a higher proportion of complex words indicates better text quality and language proficiency. Zhao and Liu [8] compared TTR and STTR of verbs, nouns, and adjectives in three corpora to study verb distribution in maritime English.

Recent research increasingly compares general English and ESP, yet comparative studies between different ESP fields remain limited. Understanding the characteristics of professional English can simplify corresponding teaching tasks.

1.2 Software WordSmith Tools, Range, and Corpus Used

WordSmith Tools is a tool for observing how words behave in text. It is useful for word frequency, word collocation, topic analysis, etc. WordSmith can be applied to all kinds of large, medium and small corpora, as well as all kinds of self-built corpora, such as JDEST, MEC and so on. It contains three text retrieval tools and three auxiliary tools, such as Concord, WordList, Keywords and so on. After Concordance is created and the proper search words are entered, Concordance will list all the examples associated with it. A list of words can also be made to directly show how often each word appears in a text file, the proportion of words in the text, how many texts each word appears in, and so on.

Range was designed by Professors Nation and Coxhead of Victoria University in New Zealand and written by Heatley. It can be used to analyze the depth and breadth of words in the text. The Range focuses on word frequency and is often used to analyze the differences between words and expressions from different sources. Moreover, Range has three basic word lists called Basewrd in general. The first two lists of baseword lists contain common words, and the third list of baseword contains academic words, which is more helpful to ESP learners. When processing single or multiple texts, Range can produce results such as frequency of occurrence, proportion of part of speech characters, and derivation forms.

This study compares Medical English Corpus (MEC), Marine Engineering English Corpus (MEE) and the Corpus of English for Science and Technology (JDEST) to discuss the features of lexical complexity in different professional English. The full name of MEE is Marine Engineering English, which is a non-coded technical English Corpus produced by Dalian Maritime University. The corpus consists of 959 representative texts with a total number of 500868 words, which is enough to support most types of vocabulary research. In 1986, under the leadership of Professor Yang Huizhong and Professor Huang Renjie of Shanghai Jiaotong University, JDEST Corpus

(Jiao Da English for Science and Technology Corpus) was put into use. The total number of words in JDEST is more than one million, with an average of 100000 words for each major, including 30% for engineering, 25% for arts and 25% for science, and 20% for biomedicine. MEC is a self-built corpus of Medical English, which contains 356 papers with a total number of 2979942 words published in the authoritative journal Lancet from 2018 to 2020.

1.3 The Present Study

This study aims to clarify the distinct characteristics of various English for Specific Purposes (ESP) by comparing the vocabulary complexity of Marine Engineering English, Medical English, and English for Science and Technology. The goal is to offer relevant suggestions for learning or professional application to learners. Lexical length, lexical sophistication, and lexical diversity have been selected as the three focal aspects of this investigation. The research aims to solve these following questions:

(1) What is the feature of word length of ESP (MEC, MEE, JDEST)?
(2) What is the feature of lexical sophistication of ESP (MEC, MEE, JDEST)?
(3) What is the feature of lexical diversity of ESP (MEC, MEE, JDEST)?

2 METHODOLOGY

2.1 Data Collection and Procedure

This study employs WordSmith and Range for both qualitative and quantitative analyses of the three corpora.

Through the WordSmith analysis of MEE, JDEST, and the self-compiled MEC, the word lists reveal the distribution of words by length. Longer words indicate higher lexical complexity, so a greater proportion of long words suggests higher lexical complexity in ESP.

Lexical sophistication is reflected by lexical frequency files and the proportion of hapax and bihapax legomena. Range is used to compile the vocabulary frequency files, which highlight the overlap of the corpus with general and academic word lists. Hapax legomena are words that appear only once in the text, while bihapax legomena appear twice. By using WordSmith's wordlist function to analyze the corpus, the frequency of each word is clearly displayed, making it easier to calculate the number of words across different corpora.

Types refer to the unique words in the corpus, and tokens represent all word forms within the corpus. The type-token ratio (TTR) is used to describe the variation and richness of vocabulary. After analyzing each corpus with WordSmith, the different TTR and standardized TTR (STTR) values are clearly presented to show lexical diversity.

3 RESULTS AND DISCUSSION

3.1 Word Length of Different ESP

Table 1. Word length in different corpora.

Length in Number of Letters	MEC		MEE		JDEST	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
1~6	2365253	67.97	369312	65.24	768144	70.53
7~12	940023	27.01	186826	33.01	305162	28.02
Over 13	174698	5.02	9908	1.75	15784	1.45
Average	4.94	-	5.60	-	5.02	-

Table 1 provides the details of word length in different corpora. In terms of word length, Medical English contains more long words, while Maritime English slightly surpasses English for Science and Technology in medium-length words. Notably, JDEST has the highest proportion of short words at 70.53%, indicating lower vocabulary complexity in English for Science and Technology. Overall, Medical English is significantly more complex than Maritime English and English for Science and Technology.

When examining individual corpora, short words (1-6 letters) account for over 60%, medium-length words (7-12 letters) for less than one-third, and long words (more than 13 letters) for less than 10% across MEC, MEE, and JDEST. This distribution pattern is consistent with their average word lengths—4.94, 5.60, and 5.02 respectively—indicating similar distributions of word lengths across the different corpora.

3.2 Lexical Sophistication of Different ESP

Frequency Profile and Condensed Frequency Profile.

Table 2. Vocabulary frequency files in different corpora.

	MEC(freq/%)		MEE(freq/%)		JDEST(freq/%)	
	Tokens/%	Types/%	Tokens/%	Types/%	Tokens/%	Types/%
Basewrd1	1486711/49.89	2980/ 3.74	350393/70.93	2630/16.30	75882/69.55	2128/22.78
Basewrd2	142805/ 4.79	1752/ 2.20	34404/ 6.96	1542/ 9.56	6874/ 6.30	992/10.62
Basewrd3	306263/10.28	2285/ 2.87	32942/ 6.67	1647/10.21	9424/ 8.64	1310/14.02
Not in the lists	1044163/35.04	72619/91.19	76246/15.43	10319/63.94	16917/15.51	4912/52.58

Table 2 indicates that in MEC, 49.89% of word tokens belong to baseword list 1, whereas in MEE and JDEST, this figure is nearly 70%, surpassing MEC's percentage. Tokens appearing simultaneously in MEC and baseword list 2 amount to 4.79%, lower than the over 6% in MEE and JDEST. For tokens in MEE and baseword list 2, the percentage is 6.96%, slightly higher than JDEST at 6.30%. Tokens in baseword list 3

and MEC account for 10.28%, which is higher than MEE (6.67%) and JDEST (8.64%). The proportion in JDEST exceeds that in MEE. In MEC, 35.04% of tokens are not in the basic vocabulary, the highest among the three. MEE and JDEST have 15.43% and 15.51% respectively of tokens not in all three baseword lists, with MEE showing a slightly lower proportion.

Regarding types, MEC has 3.74% of types appearing in both MEC and baseword list 1, much lower than MEE and JDEST. JDEST has 22.78% of types in baseword list 1, higher than MEE's 16.30%. Types in MEC and baseword list 2 are 2.20%, still lower than in MEE and JDEST. JDEST has 10.62% of types in baseword list 2, slightly more than MEE's 9.56%. Types in both MEC and baseword list 3 are only 2.87%, significantly lower than MEE's 10.21% and JDEST's 14.02%. MEE's data also surpass JDEST. The proportion of types not in the basic vocabulary is highest in MEC at 91.19%, followed by MEE at 63.94% and JDEST at 52.5%.

According to the data from the three corpora, the number and proportion of easier words (baseword list 1 and list 2) and more difficult words in different corpora are calculated and presented in the following table.

Table 3. Condensed frequency files in different corpora.

	MEC	(%)	MEE	(%)	JDEST	(%)
Easier Words	1629516	54.68	384797	77.89	82756	75.85
More Difficult Words	1350426	45.32	109188	22.11	26341	24.15
Total	2979942	100	493985	100	109097	100

The more difficult words in MEC account for 45.32%, significantly higher than in MEE (22.11%) and JDEST (24.15%). The proportion of easier words in MEE is 77.89%, slightly higher than JDEST (75.85%), both of which are much higher than in MEC (54.68%).

Based on Tables 2 and 3, it is evident that Medical English has much higher lexical sophistication than the other two types of ESP, with English for Science and Technology being slightly more sophisticated in lexis than Maritime English. This conclusion aligns with the analysis of lexical frequency profiles.

Hapax Legomena and Bihapax Legomena

Table 4. Distribution of hapax and bihapax legomena in different corpora.

	MEC	MEE	JDEST
Hapax Legomena	30103	5528	93211
Ratio of Hapax Legomena/%	43.54	36.83	69.90
Bihapax Legomena	10780	1863	13605
Ratio of Bihapax Legomena/%	15.59	12.41	10.20
Total	69125	15009	133345

Toshihiko and Shiotsu [3] believe that if some words appear more frequently in the text, there will be less cognitive impairment to deal with, and the text will be easier to understand. Table 4 indicates that the proportion of hapax legomena in JDEST is 69.90%, significantly higher than in MEC and MEE. In MEC, the proportion is 43.53%, which is also noticeably higher than MEE's 36.83%. This data suggests that the lexical complexity of English for Science and Technology is much higher than the other two types of ESP, with Medical English being more complex than Maritime English.

The proportion of bihapax legomena in MEE is 12.41%, slightly higher than in JDEST (10.20%). Both are lower than in MEC (15.59%). This data implies that the lexical sophistication of Marine Engineering English is slightly higher than that of English for Science and Technology, but both are lower than that of Medical English.

In conclusion, considering the data on hapax legomena and bihapax legomena across MEC, MEE, and JDEST, it is evident that the lexical sophistication of English for Science and Technology is higher than the other two types of ESP, and Medical English is more sophisticated than Marine Engineering English.

3.3 Lexical Diversity of Different Types of ESP

Table 5. TTRs and STTRs in different corpora.

	MEC	MEE	JDEST
Tokens	2721408	491909	1063001
Types	69125	15009	133345
Types/Tokens Ratio (TTR)	2.54	3.05	12.54
Standardized TTR (STTR)	33.62	33.48	49.86

Table 5 indicates that the TTR of MEC is 2.54, which is the lowest among the three corpora. The TTR of MEE is 3.05, which is similar to that of MEC. The TTR of JDEST is 12.54, which is much higher than those of MEC and MEE. The data clearly show that the lexical diversity of English for Science and Technology far exceeds that of Maritime English, and the lexical diversity of Maritime English is slightly greater than that of Medical English.

The corpus sizes of MEC, MEE and JDEST in this study are different, so standardized TTR is used to reduce the impact of corpus text size on the results. The standardized TTR of JDEST is 49.86, which is much higher than that of MEC and MEE. The standardized TTR values of MEC and MEE are the same. It shows that English for Science and Technology covers a wide range of vocabulary and has more different word types than do Medical English and Maritime English. The lexical diversity of Medical English is slightly more than that of Maritime English.

Combined with TTR and STTR in ME, MEE and JDEST, it can be concluded that the lexical diversity of English for Science and Technology is much higher than that of the other two types of ESP; the lexical diversity of Medical English is higher than that of Maritime English, but there is little difference between them.

4 CONCLUSION

This study analyzes the lexical complexity of different English for specific purposes with WordSmith and Range. Taking Medical English, Marine Engineering English and English for Science and Technology as examples, quantitative and qualitative methods are adopted in the study.

The analysis of word length by Wordsmith reveals that the lexical complexity of Medical English is significantly higher than Maritime English, which in turn is slightly higher than English for Science and Technology. In the aspect of lexical sophistication, focusing on lexical frequency files, Medical English is found to be much more sophisticated than English for Science and Technology, which is slightly more sophisticated than Maritime English. Data on hapax and bihapax legomena indicate that English for Science and Technology has the highest lexical sophistication, followed by Medical English and then Maritime English. Lexical diversity, measured by TTR and STTR, shows that English for Science and Technology has the highest diversity, followed by Medical English, and then Maritime English. Overall, the data suggest that the lexical complexity of Medical English is the highest, followed by English for Science and Technology, and then Maritime English.

This study is significant for English teaching and learning in professional fields such as medicine and maritime. To enhance ESP instruction, educators should develop specialized vocabulary lists tailored to each field, integrating these into practical contexts like case studies for medical students and real-life maritime scenarios. Emphasizing lexical sophistication in medical English courses and promoting lexical diversity in technical English courses can improve language proficiency. Regular assessments using tools like WordSmith and Range can help adjust teaching methods. Incorporating interactive activities, utilizing corpus-based resources, and encouraging writing practice through relevant tasks will prepare students for the linguistic demands of their professions.

REFERENCE

1. Huang, R. J., & Yang, H. Z. (1984). Preliminary analysis of computer-assisted statistical results of technical English vocabulary. Journal of Foreign Languages, (1), 46-51+28.
2. Lyu, H., & Gu, J. X. (2009). Analysis of conjunctions in discourse of maritime English based on corpus. Journal of Dalian Maritime University (Social Sciences Edition), (06), 109-112.
3. Toshihiko, S., & Shiotsu (2010). Components of L2 reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners. Cambridge: Cambridge University Press.
4. Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. Language Testing, 17(1), 65-83.
5. Wei, W. (2020). Construction and application of multimodal corpus system of medical English. Microcomputer Applications, 036(002), 75-78.

6. Wolfequintero, K., Inagaki, S., & Kim, H. Y. (1998). Second language development in writing: Measures of fluency, accuracy, & complexity. Second Language Teaching & Curriculum Center, 23(3), 423-425.
7. Zhao, X. D. (2012). Construction of maritime transportation English corpus and its quantitative study on vocabulary. Journal of Luoyang Institute of Science and Technology (Social Science Edition), (03), 13-15.
8. Zhao, X. D., & Liu, J. (2012). Quantitative study on distribution of verbs in maritime transportation English. Everyone, (12), 232-233.
9. Zhu, L. L. (2009). Characteristics of medical technology English language and teaching reform. Journal of Jilin Engineering Normal University, (10), 32-33.