https://doi.org/10.53656/ped2023-5s.12

Maritime English
Морски английски език

THE MARITIME DOMAIN-SPECIFIC CORPUS: COMPILATION AND APPLICATIONS

Yan Zhang, Shijie Liu
Shanghai Maritime University (China)

Abstract. This paper reports on the compilation of a multi-genre maritime domain-specific corpus and the research methods used to analyze it. The corpus was compiled using a combination of manual and automated methods, including web crawling and manual selection of relevant texts. Qualitative and quantitative methods, such as discourse analysis and statistical analysis, were employed to analyze the corpus. The paper describes the background, significance, text scope, principles, and process for compiling the corpus, and explores its applications, including maritime language curriculum development, standardization of maritime language genres, international maritime discourse analysis, term extraction, development of maritime domain-specific machine translation models, and the study of quantitative linguistics in the maritime domain. This study can provide a valuable resource for researchers and educators in the maritime domain and serve as a reference and inspiration for future studies in this field.

Keywords: maritime communication; corpus; development and application; term extraction; machine translation

Introduction

The compilation and research of a maritime domain-specific corpus remains a must-have in Maritime English education and in studies of communication-safety/efficiency correlation in international shipping. In the digital age, compiling a parallel corpus in the maritime domain is not only essential for China to realize strategic goals such as "Building China into a Strong Maritime Country" (2012) and advancing the *Vision for Maritime Cooperation under the Belt and Road Initiative* (2017), but it is also a must-have requirement for China to participate in international maritime governance and discourse system construction. This requires standardization of "maritime language" and research on the effectiveness of international maritime discourse behavior. To achieve these research goals, it is necessary to construct a maritime domain-specific corpus. However, a systematic literature review reveals a lack of large-scale maritime language resources, both domestically and internationally. The positioning of maritime language in corpus

construction is not clear enough, resulting in a vague scope of corpus collection and a lack of sufficient basis. Furthermore, relevant corpora are limited in size and are small-scale, self-built corpora that lack representativeness. These obstacles hinder corpus-driven maritime language research.

Based on the aforementioned issues, we will adopt a corpus-driven research approach, utilizing the "discourse analysis" framework (Gee 1999) and the three schools of genre analysis (Freedman & Medway 1994; Martin & Rose 2003; Hyland 2003; Bhatia 1993, 2004; Swales 1990), as well as "multi-goal analysis" (Tracy & Coupland 1990) to construct a "five-level language communication model" in maritime scenario. Specifically, we will utilize natural language processing methods to create the corpus and determine the context information annotation parameter framework based on the five-level language communication model. After automatic and manual annotation, analysis, noise reduction, and semantic disambiguation, we will focus on the main application scenarios of the corpus, including standardization of maritime language genres, international maritime discourse analysis, maritime dictionary compilation, construction of a maritime domain-specific term extraction model, and optimization of maritime machine translation engines in the maritime field.

1. Scope and basis of corpus data collection

1.1. ESP genre analysis

Before constructing a maritime domain-specific corpus, it is crucial to clarify the domain or position of maritime languages to determine the scope of corpus collection. In the fields of international shipping, trade, management, and diplomacy, the International Maritime Organization (IMO) of the United Nations has established the *International Convention on Standards of Training, Certification and Watchkeeping for Seafarers* (STCW), which designates English as the "lingua franca" of the international maritime industry. Therefore, the industry considers "maritime language" to be English for Specific Purposes (ESP). As the maritime language of ESP (Bocanegra-Valle 2013; Zhang&Cole 2018), the "field" includes ship-ship/shore/port communication, shipping trade, maritime law, marine engineering, and the shipbuilding industry. The "tenor" is the discourse formed between the industry and institutional personnel, and the "mode" includes both written language and spoken communication and consultation among industry personnel via VHF channels or face-to-face or remote meetings.

Research on maritime language as ESP, both domestically and internationally, can be categorized into three directions: 1) analysis of language features (Bocanegra-Valle 2013; Zhang 2016; Zhang & Zhang 2019); 2) "needs and current situation analysis" around teaching and testing (Luo & Tong 2009; Cole & Trenkner 2012), which has pointed out the urgent need for the construction of a maritime English corpus and an industry language standard framework; 3) genre analysis (Pyne &

Koester 2005; Dzeverdanovic-Pejovic 2013), which has utilized small self-built corpora to analyze maritime English discourse features (e.g., keywords, semantic load) in comparison with general English corpora.

Since Tarone et al. (1981) and Swales (1990), ESP research has been guided and framed by "genre" analysis (asturkmen 2010; Tardy 2011; Rose & Martin 2012). There are three major schools of genre analysis: 1) the "New Rhetoric" school, represented by Freedman & Medway (1994), which regards genre as a "rhetorical act"; 2) the Systemic Functional Linguistics (SFL) school, represented by Martin & Rose (2003) and Hyland (2003), which emphasizes the "register" of genre; and 3) the ESP/EAP school, represented by Bhatia (1993; 2004) and Swales (1990), which focuses on the communicative purposes and step-by-step analysis of genres. These three schools have gradually merged in ESP research (e.g., Nickerson 2000; Lockwood 2002; Pennycook 2010; Paltridge 2012). ESP research and practice emphasize that genre is a situational use of language, integrating language, communicative skills, professional abilities, and cultural factors (e.g., Nickerson 2000; Lockwood 2012; Sun 2019). Furthermore, both domestic and international "ESP genre analysis" increasingly focus on the research framework of "corpus + genre" (e.g., Flowerdew 2017; Xu 2017; Zhao et al. 2018). The study of discursive features and speech act types of ESP, particularly the investigation of metadiscourse, has received attention (e.g., Flowerdew 2008; Zhang & Wei 2019). The study of discursive features and speech act types of maritime language has been emphasized (Bocanegra-Valle 2013), and the genre framework has started to dominate maritime language education (Zhang & Cole 2018).

1.2. Scope of corpus data collection

The purpose of corpus construction plays a crucial role in determining the type and capacity of the corpus, its properties, and the extent of corpus processing (Wang & Huang 2008; Hu 2011). In this study, after identifying maritime language as an ESP genre, a "five-level language communication model" in maritime scenario is constructed (Fig.1) based on the types of maritime language genres and international maritime discourse. To construct the corpus, corpus processing tools such as alignment, coding, and segmentation (single-language corpus and bilingual parallel corpus) are employed, and the context information annotation parameter framework is determined based on the "five-level language communication model".

The research focuses on "maritime language" as the object of study, which comprises two categories: the genre chain/colony of maritime language genres and international maritime discourse. The genre chain/colony of maritime language genres (Bhatia 2004; Swales 2004) is characterized by both spoken and written language. The spoken language encompasses the "ship-shore/ship communication genre group," which is formed based on industry communication scenarios. This includes ship-VTS communication, on-board communication, pilotage communication, Port State Control (PSC) communication, and ship-to-

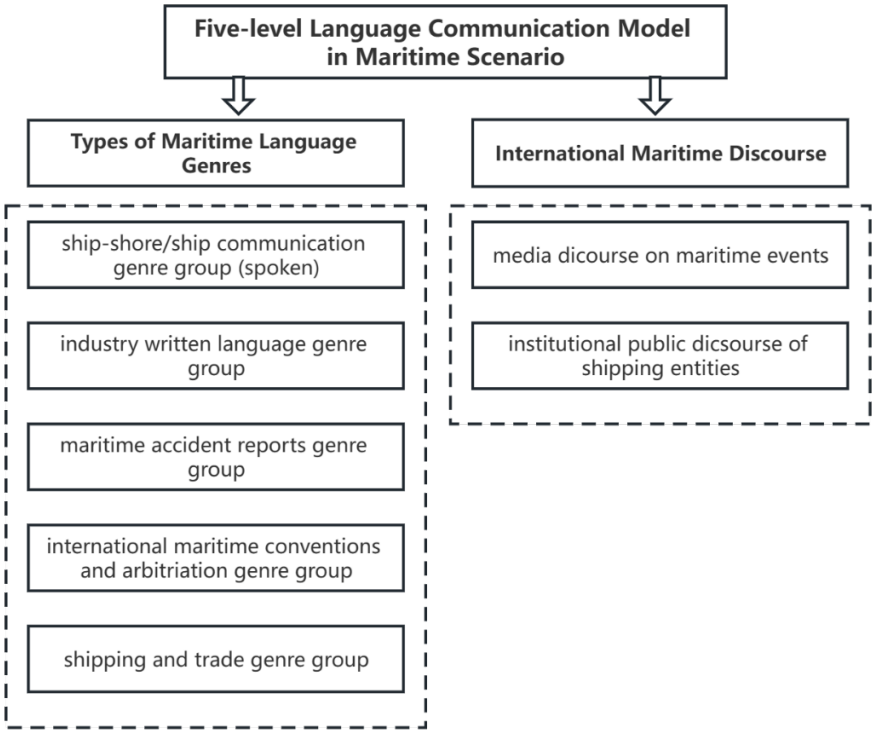

Figure 1. Five-level language communication model in maritime scenario[1]

ship communication (collision avoidance, salvage, etc.). The written language is categorized into: 1) industry written language genre group, which includes navigation written language, marine engineering written language, and other maritime written language for different positions; 2) maritime accident reports; 3) English-Chinese parallel texts of international maritime conventions and maritime arbitration; and 4) shipping and trade genre group, comprising bill of lading corpus and lease corpus. The research aims to explore the standardization of maritime language and the correlation between maritime communication and maritime safety and efficiency based on the aforementioned research objects.

International maritime discourse encompasses: 1) media discourse on maritime events, including English-language editorials, news reports, and feature stories from mainstream media in China, the United States, the United Kingdom, Australia, and Canada over the past decade, with a focus on international maritime events; and 2) institutional public discourse, such as proposals and meetings of the IMO, discussions of maritime and oceanic think tanks, annual reports, corporate social responsibility (CSR) reports, and sustainable development reports of international maritime non-

governmental organizations, and the top 100 global shipping companies according to the Alphaliner TOP 100 data[2]. Analyzing the above discourse resources, the study identifies the main themes of international maritime discourse, discourse concept dissemination, and discourse system structure. Additionally, it identifies different voices and discourse motivation semantic models, image construction, and crisis handling in the international maritime discourse community. Furthermore, the corpora constructed based on the two categories can be used to optimize the term extraction model, improve the effectiveness of machine translation engines, and conduct quantitative linguistic research in the maritime field.

2. The value of construction of multi-genre maritime domain-specific corpus

2.1. Academic value

Constructing a maritime domain-specific corpus holds significant academic value for research in the maritime field. This study's academic value is reflected in three aspects: Firstly, by utilizing large-scale and systematic maritime language corpus data, the research methodology of "corpus + genre" is employed to create a standardized description and parameter framework of maritime language. This framework clarifies how the "effectiveness" of maritime communication directly impacts international maritime safety and efficiency. This expands and improves the theoretical framework and methodology of linguistic ESP, which is instrumental in leading the formulation of international maritime lingua franca standards. Secondly, this study expands the research of quantitative linguistics in the maritime domain, examining whether the laws summarized in the general domain differ in the maritime field, and attempts to explore the unique characteristics of maritime language. Thirdly, it develops a "critical + positive" discourse analysis framework, systematically constructing the discourse concept dissemination and discourse system related to China's national strategy of "Maritime Silk Road" and "strengthening the system of maritime rights and interests and participating deeply in international maritime governance."

2.2. Application value

Constructing a maritime domain-specific corpus can provide many values for research and applications. Firstly, by collecting and organizing a large amount of maritime domain text, a knowledge base of the maritime field can be established, containing a wealth of information such as terminology, concepts, and entities. This information can be utilized in various maritime applications, such as maritime information retrieval and maritime knowledge graph construction. Secondly, the maritime domain-specific corpus can be employed to train and evaluate natural language processing models, such as text classification, terminology extraction, named entity recognition, and machine translation engine optimization. By utilizing the maritime domain-specific corpus, the performance and accuracy of these models in the maritime field can be improved, better supporting maritime applications and research. Thirdly, the maritime domain-specific corpus can also be used to analyze

and study trends and development directions in the maritime field. By analyzing maritime domain texts, information such as hot issues, key technologies, and development trends in the maritime field can be understood, providing reference for maritime-related decision-making and planning. Finally, the corpus can also support maritime education and training. By using the maritime domain corpus, more practical education and training content can be provided for maritime students and professionals, improving their professional level and ability in the maritime field.

3. The process of building multi-genre maritime domain-specific corpus

3.1. Corpus data pre-processing

3.1.1. Written language text recognition and proofreading, noise reduction

To convert the scanned PDF document into an editable *.doc(x) document, ABBYY FineReader 15, an optical recognition software, is utilized, providing the basis for proofreading. Manual intervention is then performed to manually proofread and reduce the noise of the recognized *.doc(x) documents. Based on the corpus text processing experience of previous researchers (e.g., Hu 2011; Li & Hu 2021), proofreading in the process of building a parallel corpus mainly involves correcting typos, garbled characters, removing extra spaces, line breaks, etc. Denoising mainly involves removing content or formats that do not conform to the corpus building standards, such as headers, page numbers, annotations, images, etc. Considering the large volume of text, a dedicated text processing team is formed. Therefore, when team members collaborate on proofreading and denoising, unified principles and standards are necessary to adopt. The edited text is saved in plain text (*.txt) format and encoded in UTF-8 format for subsequent storage and processing.

3.1.2. Processing of audio files

For annotating audio corpus (in English or Chinese), Praat software, a computer speech analysis software developed by two scholars, Paul Boersma and David Weenink, from the University of Amsterdam in the Netherlands, is planned to be used. Praat software can visualize phonetic attributes of the corpus, such as speech, pitch, intonation, and stress, through spectrograms and pitch contours, and perform analysis and annotation based on these visualizations. The software annotation is designed with multiple layers, and the specific number of layers can be set by the user. Within the same annotation layer, boundaries can be freely divided by setting, as described by Bei & Xiang (Bei & Xiang 2016, p.46). The hierarchical annotation function of the software is not only suitable for phonetic features but can also be applied to non-fluent features, meeting research needs.

Before using Praat to annotate the corpus, preprocessing of the audio corpus is necessary, including clipping, format conversion, precise alignment, dual-track file creation, and AI pre-transcription (iFlytek Tingjian). Clipping is utilized to extract the initial research corpus from the resource library. The initial corpus may be in different audio or video formats, which can be uniformly converted into *wav

format audio using Format Factory software. The initial format corpus files are retained for future research extensions. When using the online transcription tool "iFlytek Tingjian," the accuracy of transcription may be significantly reduced when encountering severe speech accents or significant environmental background noise. However, manually transcribing based on the machine-transcribed draft can still greatly improve the efficiency of manual operation.

3.2. Segmentation and annotation

Word segmentation refers to the separation of connected characters into separate symbols. Annotation refers to marking the attributes of text in the corpus using tags. For written language text or text transcribed from audio files, different levels of annotation are performed based on the research purpose of standardizing maritime language genre and analyzing international maritime discourse. These may include part-of-speech tagging, prosodic annotation, turn-taking annotation, and metadata annotation. For maritime convention texts with bilingual versions, additional annotation of translation strategies and techniques is required, using closed tags and involving vocabulary, structure, syntax, and other levels.

Part-of-speech tagging is the process of assigning a part of speech to each word in a text based on contextual information. Metadata, on the other hand, refers to information that describes information. Adding metadata annotations to a corpus can provide retrieval conditions and a query basis for later retrieval, statistics, and analysis (Li & Hu 2021, p. 85). According to Liang et al. (Liang et al. 2010, p. 37), using metadata to retrieve a corpus is an advanced application of the corpus. Metadata annotation can provide query conditions and a basis for corpus retrieval and analysis, mainly involving the annotation of prosodic and internal structural information of the text. The annotation method involves annotating the beginning and end of each prosodic or internal structure in the text with a unified format (Feng & Wu 2017). Metadata is typically annotated at the beginning of each corpus text, indicating the external information of the text, such as the title, author, time, source, and text classification. The annotation of text structure information and grammar information mainly includes the internal information of the text, such as title, chapter, paragraph, sentence, and part-of-speech tagging.

3.3. Text alignment

Bilingual texts of maritime conventions and arbitration were aligned at the paragraph level using alignment software ABBYY Aligner 2.0 and manual adjustment. Sentence-level alignment between English and Chinese was achieved based on the paragraph alignment, and the aligned texts were exported in *.txt or *.xml format. To maximize the utilization of bilingual texts of maritime conventions and arbitration in the past decade (2010 – 2022), corpus alignment was necessary to align the original texts and translations of relevant texts. Corpus alignment involves aligning bilingual or multilingual texts at the paragraph, sentence, or word level and establishing a one-to-one correspondence between the source language and the

target language at the alignment level (Wang et al. 2019, p.32). This project adopted sentence-level corpus alignment, considering the degree of corpus utilization. Common corpus alignment tools/platforms include ABBYY Aligner, Tmxmall online alignment, and the memory module of computer-assisted translation (CAT) tools. After comparison and selection, ABBYY Aligner 2.0 was chosen as the tool for corpus alignment. After alignment, a translation memory file (*.tmx) was obtained and exported as *.txt or *.xml format using HeartSome TMX Editor.

3.4. Dataset annotation

To fully utilize the constructed corpus, such as extracting terminology in the maritime field, term annotation is required for the maritime language dataset. Commonly used annotation methods for term annotation tasks include BIO, IOB, IOB2, and BMEWO. Among them, the BIO (Beginning, Inside, Outside) annotation code decomposes each term into several words and annotates each word, which is divided into three types: B (Beginning), I (Inside), and O (Outside). B represents the beginning of a term, I represents the middle part of a term, and O represents not belonging to any term. The BMEWO (Beginning, Middle, End, Whole, Other) annotation code is an extension of BIO, which divides terms into five types: B (Beginning), M (Middle), E (End), W (Whole), and O (Other). B represents the beginning of a term, M represents the middle part of a term, E represents the end part of a term, W represents a complete term, and O represents not belonging to any term. These annotation methods are suitable for various text forms, including horizontally spread and vertically listed text. The text used in this study is vertically listed after processing, where each character is listed separately as a line or column. After observing the maritime language text set, it was found that there is a problem of nested terms in the maritime field, where a shorter term may appear in a longer term, and vice versa.

This study selected representative maritime texts to construct a dataset of about 500,000 tokens and manually annotated terms using the BIO annotation method. In the construction of the term extraction model, it involves data preprocessing (word segmentation, part-of-speech tagging, entity tagging, etc.), feature extraction (word embedding, context information, entity types, etc.), model selection (choosing models suitable for the task and dataset, including CNN, RNN, Attention, etc.), model training (using a large amount of annotated data to train the model and adjusting the model parameters and hyperparameters to obtain better performance), and other steps. Deep learning-based models are used for term extraction, taking into account complex term nesting situations in the maritime field.

4. Main applications of multi-genre maritime domain-specific corpus

4.1. Corpus-based dictionary compilation and word list development

Corpus has become the primary source of lexicographers in compiling dictionaries, providing them with raw data for analyzing word meanings and usage (Rundell 2009a;

2009b). Large-scale corpora have become a prerequisite and main tool for dictionary compilation. Therefore, a maritime language corpus-based dictionary compilation platform can be built, which provides menu items such as List (frequency table), Collocates (high-frequency collocations), and KWIC (keyword in context) and provides vocabulary analysis functions, including frequency tables, word retrieval, context, retrieval result sorting, vocabulary collocation, synonyms, and antonyms. Overall, bilingual corpus-based dictionaries can ensure that all word meanings and syntactic information are verified by real corpora, ensuring the reliability and accuracy of information. Currently, the Shanghai Maritime University Maritime English Dictionary Retrieval Platform[3] has been launched, and all core vocabulary and example sentences are extracted and retrieved based on real corpora.

There are two ways to develop a word list: 1) terminology extraction software can be used on the maritime language corpus (taking the "Maritime Convention English-Chinese Parallel Corpus" as an example) to extract terms, and comprehensive measures such as precision, recall, and *F1* measure should be balanced (Vivaldi & Rodríguez 2007). The recall rate and *F1* measure should be calculated under the premise of strictly formulating a gold standard (Wang & Liu 2022, p. 51), and the final word list development needs to be manually screened and determined; 2) with the help of existing corpus retrieval tools, such as AntConc and WordSmith Tools 8.0, keyword lists can be generated based on specific texts and compared and screened with relevant reference keyword lists to determine the final keyword list.

4.2. Promoting the standardization of maritime language genres

In the field of standardization research in maritime language genres, a "corpus + genre" micro-quantitative analysis is used to locate the goal-driven, normative nature of maritime English as an international industry lingua franca and its correlation with maritime safety. The focus can be on the following six aspects: 1)analysis of written language patterns (Bhatia 1993; 2004); normative conversation analysis of natural dialogue "turn-taking", "adjacency pairs", and "sequential organization" (Drew & Heritage 1992; Schegloff 1992) and their indications for the hierarchical contexts of "macro discourse", "social behavior", and "professional behavior" (Handford 2010); 2)maritime language terminology and its distribution; comparison of the language actually used in the maritime industry with the International Maritime Standard Communication Phrases (SMCP). 3)cross-cultural scripts reflected in "small talk" and "code switching" in maritime language; 4) micro-language pattern parameters such as word clusters, vocabulary growth, word frequency distribution, semantic word distribution, and grammar diversity distribution between different genre sub-corpora in maritime language and with general English language corpora (such as Brown Corpus of Standard American English, Vienna-Oxford International Corpus of English); 5)interpersonal meaning language expressions in maritime language (pronouns, vague language, obligation modality, etc.); 6) high-frequency "speech acts" (Ruhleman & Aijmer 2015) in maritime language, increasing the amount

of text, using R/Python to focus on logistic linear/non-linear regression analysis, establishing a visual correlation between maritime language and maritime safety through the "risk of speech act misunderstanding" spectrum.

By conducting a micro-quantitative analysis of these six aspects, researchers can gain a deeper understanding of the normative nature of maritime English and its correlation with maritime safety, which can inform the development of standardized language practices and training programs for the maritime industry.

4.3. Promoting the analysis of international maritime discourse

In the realm of international maritime discourse research, a comprehensive framework comprising critical discourse analysis and positive discourse analysis, along with text mining techniques, can be employed to address three key types of issues based on the corpus.

Firstly, by utilizing keyword and collocation corpus mining techniques on the "Maritime Events Media" corpus from the past decade, themes and features of international maritime media discourse, particularly those pertaining to China or Chinese involvement in maritime events, can be identified. For instance, Zhang Yan's (2008) framework can be utilized to discern the discourse motivations of various discourse subjects and distinguish different international positions. Correspondingly, China should adopt a suitable "discourse feedback" mode to achieve moderate "discourse control" and truly shape the "agenda setting" and "discourse guidance" in the international maritime domain.

Secondly, the establishment and dissemination of the concepts of China's "Maritime Power Strategy" and "Maritime Silk Road" discourse, as well as the pursuit of marine governance rights and the setting of international maritime order topics in a complex international environment, are important topics for research.

Thirdly, research topics related to marine environmental protection, the shipping industry's role in achieving global sustainable development goals, responding to global crises such as the COVID-19 pandemic (Zhang & Sun 2021; Sun & Zhang 2022), corporate social responsibility (CSR) (Vishwanathan et al. 2020), non-governmental organizations (NGOs), and shipping companies are also significant.

4.4. Identifying the link between maritime language and maritime safety

The relationship between maritime communication and maritime safety is crucial. In practical maritime work, maritime practitioners engage in extensive language communication to ensure safe and smooth operation of maritime transportation. Correct, clear, and standardized language communication can prevent misunderstandings and errors, and reduce accidents. The higher the standardization of maritime language, the easier it is for maritime practitioners to understand and communicate with each other accurately, thereby improving maritime safety and efficiency. Research on the characteristics of maritime discourse based on corpora can help us better understand the characteristics and norms of maritime language, thereby avoiding or reducing maritime accidents.

In recent years, research on the characteristics of maritime discourse based on corpora has gained significant attention. These studies use methods such as compiling ship-shore dialogue corpora, transcribing and analyzing voyage data recorders to quantify and analyze the compliance rate of standard protocols in real routine ship-shore communications (Jurkovič 2022), information flow in bridge team communications (John et al. 2013), and speech acts in verbal communication (John et al. 2019). The aim is to evaluate the efficiency and safety of maritime communication and improve the communication quality in the maritime field. These studies provide valuable references for ship-to-ship/ship-to-shore communication in the maritime field and help improve the efficiency and safety of maritime communication.

Therefore, research on the characteristics of maritime discourse based on corpora has a close relationship with maritime language and maritime accidents, and it can help improve the safety and sustainable development of the maritime field. In the future, based on the established corpus of maritime accident reports, multidimensional annotation and deep text mining can be conducted on maritime accident reports. This approach can help identify the correlation between language communication and maritime safety, thereby improving the effectiveness of maritime language communication and reducing international maritime safety accidents caused by non-standardized maritime language. Additionally, by improving the quality and standard of maritime language communication, it can comprehensively enhance the soft environment in shipping enterprises/institutions.

4.5. Training to form a maritime domain-specific term extraction model

Term extraction is a crucial knowledge unit in specific professional fields, providing benefits for several terminology tasks and downstream tasks such as information retrieval, machine translation, topic detection, sentiment analysis, etc. (Tran et al. 2023, p.1). The maritime field involves a wide range of subdivided fields, complex features, and a lack of large-scale mature corpora, making term extraction difficult. Timely mastering maritime terms not only helps to dynamically grasp the development direction of the maritime field, reveal the core knowledge and research hotspots, but also promotes the standardization research of maritime language under the vision of maritime power strategy.

To improve the term extraction performance and reduce the manual annotation cost of the dataset, deep learning-based term extraction research in the maritime field is an important issue that requires attention. Firstly, a corpus collection with representative maritime language should be constructed and manually annotated to establish a labeled dataset and a golden standard. Secondly, a large amount of monolingual corpus in this field is used to train and optimize other terminology extraction models to improve the model's generalization ability and accuracy. Thirdly, the model is trained, optimized, and tested based on the manually annotated dataset to improve the model's adaptability and accuracy in the maritime

field. Finally, the performance of the model is tested using test texts in this field and other existing terminology extraction models in the industry to evaluate the comprehensive performance of the model (such as precision, recall, *F1* measure).

4.6. Optimizing the effectiveness of maritime domain-specific machine translation engine

The increasing frequency and complexity of communication and cooperation in the maritime field due to globalization have led to a growing interest in the application of machine translation in this field. However, due to the high professionalism and specificity of the maritime field, traditional machine translation models often fail to meet the needs of the field, including professional terminology, abbreviations, and specific contexts, which can affect the quality and efficiency of translation. To address this issue, transfer learning has emerged as a powerful method that can fully leverage existing corpora, avoid training models from scratch, save time and resources, and optimize the performance of machine translation engines by leveraging existing maritime corpora.

The multi-genre maritime domain-specific corpus includes a bilingual corpus sub-library composed of shipping reports and maritime arbitration, which can be combined with monolingual corpora, terminology intervention, and vocabulary sharing to conduct semi-supervised neural machine translation training. Additionally, basic machine translation models such as Google's Seq2Seq model can be used to train general corpora and then applied to maritime corpora for transfer learning. By fine-tuning the model parameters, using multi-task learning or domain adaptation methods, the model can adapt to the characteristics of the maritime field and improve the translation effect of maritime field texts.

4.7. Contributing to the study of quantitative linguistics in the maritime domain

The multi-genre maritime corpus is a valuable research resource that provides data support for language research in the maritime field and expands the application of quantitative linguistics in vertical fields. Maritime language is a unique form of language with distinctive language features and expression methods. By conducting quantitative analysis of maritime language, we can explore the quantitative linguistic features of maritime language and understand its uniqueness. The distributional law (Zipf's Law), functional law (Menzerath-Altmann Law), development law (Piotrowski-Altmann Law), and laws of word length distribution, word length and word frequency correlation, word length and polysemy correlation, polysemy and synonymy correlation, synonymy and word length correlation, frequency and multilingualism correlation, text block law, and the relationship between component order, length, and complexity are all directions that can be studied (Liu 2017).

Maritime language includes many professional terminologies and technical vocabularies, and the length and frequency distribution of these vocabularies

may differ from those of general language. In addition, grammar is an essential component of language, which includes component order, grammatical structure, grammatical complexity, and other aspects. The grammatical structure and component order in maritime language may differ from those of general language, and these differences may be related to the professional characteristics of the maritime field. Therefore, it is crucial to explore the unique features of maritime language based on the established corpus and examine whether there are differences in the laws of quantitative linguistics between general and maritime vertical fields.

Conclusions and summary

This article introduces the foundation of maritime language as English for Specific Purposes (ESP) and proposes a "five-level language communication model" in maritime scenario based on the genre types of maritime language and international maritime discourse. The model is constructed using the "discourse analysis" framework and the three major schools of genre analysis (Gee 1999), as well as "multi-objective analysis" (Tracy & Coupland 1990). Additionally, a multi-genre maritime domain-specific corpus is developed, and the research value, constructing process, and main application scenarios of the corpus are elaborated. The corpus has numerous potential applications, including corpus-based dictionary compilation, term extraction model construction, and optimization of the maritime domain-specific machine translation engine. The establishment of this multi-genre maritime domain-specific corpus has significant research and application value, providing support for teaching, research, and practice in the maritime field. The article concludes by emphasizing the need to continuously improve the corpus's quality, application effects, and annotation dimensions to contribute to the development of the maritime field.

NOTES

1. This model was informed by the "discourse analysis" framework (Gee 1999) and the three schools of genre analysis (Freedman & Medway 1994; Martin & Rose 2003; Hyland 2003; Bhatia 1993, 2004; Swales 1990), along with "multi-goal analysis" (Tracy & Coupland 1990). Based on this foundation, Prof. Yan Zhang and her team members constructed the Five-level Language Communication Model in Maritime Scenario.
2. ALPHALINER. Alphaliner TOP 100. Web site. Available from: https://alphaliner.axsmarine.com/PublicTop100/. [Viewed 2023-04-28].
3. SHANGHAI MARITIME UNIVERSITY. *Shanghai Maritime University Maritime English Dictionary Retrieval Platform.* Web site. Available from: http://dict.shmtu.edu.cn/yizhe/dic/home. [Viewed 2023-04-28].

Acknowledgement

The work was supported and funded by the National Social Science Fund of China, grant number 21BYY017, and the 2022 Top-notch Innovative Talents Cultivation Program for Graduate Students of Shanghai Maritime University, grant number 2022YBR020.

REFERENCES

BASTURKMEN, H., 2010. *Developing Courses in English for Specific Purposes*. Basingstoke: Palgrave Macmillan. ISBN 978-023-0227-97-2. Available from: https://doi.org/10.1057/9780230290518.

BEI, X.; XIANG, N., 2016. *Principles of Experimental Phonetics and Operation of Praat Software*. Changsha: Hunan Normal University Press. ISBN 978-756-4824-95-2.

BHATIA, V. K., 1993. *Analyzing Genre: Language Use in Professional Settings*. London: Longman. ISBN 978-058-2085-24-4.

BHATIA, V. K., 2004. *Worlds of Written Discourse: A Genre-based View*. London: Continuum. ISBN 978-147-2522-63-4.

BOCANEGRA-VALLE, A., 2013. Maritime English. In: C. A. CHAPELLE (ed.), *The Encyclopedia of Applied Linguistics*, pp. 3570 – 3583. West Sussex: Wiley-Blackwell. ISBN 978-140-5194-73-0.

COLE, C. W.; TRENKNER, P., 2009. The Yardstick for Maritime English STCW assessment purposes. *IAMU Journal,* vol. 6, no. 1, pp.13 – 28. ISSN 1302-678X.

COLE, C. W.; TRENKNER, P., 2012. The STCW Manila Amendments and their impact on Maritime English. *Constanta Maritime University's Annals*, no. 17, pp. 239 – 244. ISSN 1582-3601.

DREW, P.; HERITAGE, J. (eds.), 1992. *Talk at Work: Interaction in Institutional Settings*. Cambridge: Cambridge University Press. ISBN 978-052-1376-33-4.

DŽEVERDANOVIĆ-PEJOVIĆ, M., 2013. Discourse of VHF communication at sea and the intercultural aspect. *International Journal for Traffic & Transport Engineering*, vol. 3, no. 4, pp. 377 – 396. Available from: https://doi.org/10.7708/ijtte.2013.3(4).03.

FENG, X., et al., 2017. Construction of MedAca medical academic English corpus. *Corpus Linguistics*, vol. 4, no. 2, pp. 107 – 116.

FLOWERDEW, J., 2017. Corpus-based approaches to language description for specialized academic writing. *Language Teaching*, vol. 50, no. 1, pp. 90 – 106. ISSN 0261-4448. Available from: https://doi.org/10.1017/S0261444814000378.

FLOWERDEW, L., 2008. *Corpus-based Analyses of the Problem-Solution Pattern*. Amsterdam: John Benjamins. ISBN 978-902-7223-03-6.

FREEDMAN, A.; MEDWAY, P. (eds.), 1994. *Genre and the New Rhetoric*. London & New York: Taylor & Francis. ISBN 978-074-8402-57-1.

GEE, J. P., 1999. *An Introduction to Discourse Analysis: Theory and Method*. London & New York: Routledge. ISBN 978-041-5725-56-9.

HANDFORD, M., 2010. What can a corpus tell us about specialist genres? In: A. O'KEEFFE & M. J. MCCARTHY (eds.). *The Routledge Handbook of Corpus Linguistics*, pp. 255 – 269. London & New York: Routledge. ISBN 978-036-7076-38-2.

HU, K., 2011. *An Introduction to Corpus Translation Studies*. Shanghai: Shanghai Jiao Tong University Press. ISBN 978-731-3071-40-8.

HYLAND, K., 2003. Genre-based pedagogies: A social response to process. *Journal of Second Language Writing*, no. 12, pp. 17 – 29. ISSN 1873-1422. Available from: https://doi.org/10.1016/S1060-3743(02)00124-8.

JOHN, P., et al., 2013. Information density in bridge team communication and miscommunication—A quantitative approach to evaluate maritime communication. *WMU Journal of Maritime Affairs*, vol. 12, no. 2, pp. 229 – 244. ISSN 1651-436X, 1654-1642. Available from: https://doi.org/10.1007/s13437-013-0043-8.

JOHN, P.; BROOKS, B; SCHRIEVER, U., 2019. Speech acts in professional maritime discourse: A pragmatic risk analysis of bridge team communication directives and commissives in full-mission simulation. *Journal of Pragmatics*, vol. 140, pp. 12 – 21. ISSN 0378-2166. Available from: https://doi.org/10.1016/j.pragma.2018.11.013.

JURKOVIĆ, V., 2022. Authentic routine ship-shore communication in the Northern Adriatic Sea area – A corpus analysis of discourse features. *English for Specific Purposes*, no. 68, pp. 47 – 59. ISSN 0889-4906. Available from: https://doi.org/10.1016/j.esp.2022.06.002.

LI, X.; HU, K., 2021. The multilingual parallel corpus of Xi Jinping: *The Governance of China*: Compilation and applications. *Technology Enhanced Foreign Language Education (TEFLE)*, no. 3, pp. 83 – 88. ISSN 1001-5795.

LIANG, M.; LI, W.; XU, J., 2010. *Corpus Application Tutorial*. Beijing: Foreign Language Teaching and Research Press. ISBN 978-756-0098-44-9.

LIU, H., 2017. *An Introduction to Quantitative Linguistics*. Beijing: Commercial Press. ISBN 978-710-0150-21-7.

LOCKWOOD, J., 2002. *Language programme training design and evaluation processes in Hong Kong workplaces*. Doctoral dissertation. Hong Kong: University of Hong Kong. Available from: The Education University of Hong Kong, https://bibliography.lib.eduhk.hk/en/bibs/246667f8.

LOCKWOOD, J., 2012. Developing an English for specific purpose curriculum for Asia call centries: How theory can inform practice. *English for Specific Purposes*, no. 31, pp. 14 – 24. ISSN 0889-4906. Available from: https://doi.org/10.1016/j.esp.2011.05.002.

LUO, W.; TONG, D., 2009. Teaching methods and research approaches for specialized English: Taking maritime English as an example. *Foreign Language World*, vol. 130, no. 1, pp. 86 – 96. ISSN 1004-5112.

MARTIN, J. R.; ROSE, D., 2003. *Working with Discourse: Meaning beyond the Clause*. London: Continuum. ISBN 978-082-6488-50-3.

NICKERSON, C., 2000. *Playing the Corporate Language Game: An Investigation of the Genres and Discourse Strategies in English Used by Dutch Writers Working in Multinational Corporations*. Amsterdam: Rodopi Bv Editions. ISBN 978-904-2007-30-7.

PALTRIDGE, B., 2012. Genre and English for specific purposes. In: B. PALTRIDGE & S. STARFIELD (eds.). *The Handbook of English for Specific Purposes*, pp. 347 – 366. West Sussex, UK: Wiley Blackwell. ISBN 978-111-8941-55-3.

PENNYCOOK, A., 2010. *Language as a Local Practice*. London: Routledge. ISBN 978-041-5547-50-5.

PYNE, R.; KOESTER, T., 2005. Methods and means of analysis of crew communication in the maritime domain. *The Archives of Transport*, vol. 17, no. 3 – 4, pp. 1-16. ISSN 0866-9546.

ROSE, D.; MARTIN, J. R., 2012. *Learning to Write/Reading to Learn: Genre, Knowledge and Pedagogy in the Sydney School*. London: Equinox. ISBN 978-184-5531-44-7.

RUHLEMANN, C.; AIJMER, K., 2015. *Corpus Pragmatics. A Handbook*. Cambridge: Cambridge UP. ISBN 978-110-7015-04-3.

RUNDELL, M.; XIA, L.; ZHU, D., 2009. The latest developments and future trends of corpus lexicography (Part 1) – Explicit use of corpus data in learning dictionaries. *Lexicographic Research*, vol. 171, no. 3, pp. 71 – 78. ISSN 1000-6125.

RUNDELL, M.; XIA, L.; ZHU, D., 2009. The latest developments and future trends of corpus lexicography (Part 2) – Explicit use of corpus data in learning dictionaries. *Lexicographic Research*, vol. 171, no. 4, pp. 81 – 91. ISSN 1000-6125.

SCHEGLOFF, E. A., 1992. Introduction. In: H. SACKS, G. JEFFERSON & E. A. SCHEGLOFF (eds.). *Lectures on Conversation*. Oxford: Blackwell. ISBN 978-155-7867-05-6.

SUN, Y., 2019. Implementing the Requirements of National Standards and Improving the Training of Foreign Language and Literature Majors. *Foreign Languages in China*, vol. 16, no. 5, pp. 36 – 42. ISSN 1672-9382.

SUN, Z.; ZHANG, Y., 2022. Strategic crisis response of shipping industry in the post COVID-19 era: A case of the top 10 shipping lines. *Journal of Marine Science and Engineering*, vol. 10, no. 5. ISSN 2077-1312. Available from: https://doi.org/10.3390/jmse10050635.

SWALES, J. M., 2004. *Research Genres: Explorations and Applications*. Cambridge, UK: Cambridge University Press. ISBN 978-052-1825-94-8.

SWALES, J. M., 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press. ISBN 978-052-1338-13-4.

TARDY, C. M., 2011. ESP and multi-method approaches to genre analysis. In: D. BELCHER, A. JOHNS & B. PALTRIDGE (eds.). *New Directions in English for Specific Purposes Research,* pp.143 – 173. Ann Arbor, MI: University of Michigan Press. ISBN 978-047-2034-60-4.

TARONE, E., et al., 1981. On the use of the passive in two astrophysics journal papers. *The ESP Journal*, no.1, pp. 123 – 140. ISSN 0272-2380. Available from: https://doi.org/10.1016/0272-2380(81)90004-4.

TRACY, K.; COUPLAND, N. (eds.), 1990. *Multiple Goals in Discourse*. Bristol, UK: Multilingual Matters. ISBN 978-185-3590-99-3.

TRAN, H. T. H., et al., 2023. The recent advances in automatic term extraction: A survey. *arXiv preprint arXiv: 2301.06767*. ISSN 2331-8422.

VISHWANATHAN, P., et al., 2020. Strategic CSR: A concept building meta-analysis. *Journal of Management Studies*, vol. 57, no. 2, pp. 314 – 350. ISSN 2302-8122. Available from: https://doi.org/10.1111/joms.12514.

VIVALDI, J.; RODRÍGUEZ, H., 2007. Evaluation of terms and term extraction systems: A practical approach. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, vol. 13, no. 2, pp. 225 – 248. ISSN 0929-9971. Available from: https://doi.org/10.1075/term.13.2.06viv.

WANG, H.; CHEN, N.; YE, M., 2019. *100 Questions on Translation Technology*. Beijing: Science Press. ISBN 978-703-0644-40-4.

WANG, H.; LIU, S., 2022. On the evaluation framework of terminology extraction software. *China Terminology*, vol. 24, no. 1, pp. 45 – 54. ISSN 1673-8578.

Xu, J., 2017. Research and compilation of medical academic English dictionary from the perspective of genre phrases. *Foreign Languages and Their Teaching*, vol. 297, no. 6, pp. 52 – 60. ISSN 1004-6038.

ZHANG, X., 2016. Metaphorical and metonymic expressions of maritime terminology in English and Chinese. *Journal of Shanghai Maritime University*, vol. 37, no. 2, pp. 94 – 102. ISSN 1672-9498.

ZHANG, X.; ZHANG, F., 2019. Study on the formation and characteristics of yangjingbang-style maritime English morphology based on spoken corpus. *Journal of Dalian Maritime University (Social Science Edition)*, vol. 18, no. 2, pp. 107 – 116. ISSN 1671-7031.

ZHANG, Y., 2008. Study on dynamic genre from the perspective of topology. *Contemporary Rhetoric*, vol. 145, no. 1, pp. 19 – 24. ISSN 1674-8026.

ZHANG, Y.; COLE, C., 2018. Maritime English as a code-tailored ESP: Genre-based curriculum development as a way out. *Ibérica*, no. 35, pp. 145 – 170. ISSN 0211-0776.

ZHANG, Y.; SUN, Z., 2021. The coevolutionary process of maritime management of shipping industry in the context of the COVID-19 pandemic. *Journal of Marine Science and Engineering*, vol. 9, no. 11. ISSN 2077-1312 Available from: https://doi.org/10.3390/jmse9111293.

ZHANG, Y.; WEI, N., 2019. A study on the characteristic meaning of nominalization in academic English by Chinese scholars. *Foreign Languages and Their Teaching*, vol. 308, no. 5, pp. 58-73, p. 149. ISSN 1004-6038.

ZHAO, Y., et al., 2018. A comparative study on English abstracts of empirical articles in Chinese and foreign academic journals. *Foreign Languages and Their Teaching*, vol. 298, no. 1, pp. 61-71, pp. 147-148. ISSN 1004-6038.

✉ **Mr. Yan Zhang**
Mr. Shijie Liu, PhD student
ORCID iD: 0009-0006-2674-6494
Web of Science Researcher ID: HNR-4916-2023
Shanghai Maritime University
Shanghai, China
E-mail: chriszy924@163.com
E-mail: henryliushijie@163.com