MARITIME ENGLISH REGISTER: A CORPUS-BASED STUDY OF ITS SPOKEN AND WRITTEN SUBVARIETIES

Jana KEGALJ
University of Rijeka
Sandra TOMINAC COSLOVICH
University of Rijeka

Abstract: *Maritime English, a language variety in the domain of occupational English, is considered the official language of communication in shipping. This study aims to provide a corpus-based description of language characteristics of subvarieties of Maritime English. This description will serve as the basis to analyze the distinctive use of linguistic features in different subvarieties. The study was conducted on four corpora: a spoken corpus of Maritime Communications, and three written corpora: a corpus of Maritime Law, a corpus of Marine Engineering English and a corpus of Maritime Academic English. Corpus data, e.g. parts of speech frequencies, noun and verb frequencies, wordlists, keywords and key terms lists, and n-grams, were used to identify some of the recurring features or patterns. These were then further analyzed qualitatively to explore their functional and pragmatic aspects. The study is an attempt to provide insight into generic features and patterns of variation in maritime discourse and compare different subvarieties within the Maritime English register. The results may have possible pedagogical implications for teaching Maritime English by providing insights into how various disciplines manipulate linguistic resources to achieve different functional goals. This could enable course designers to create domain-specific syllabi tailored to the actual needs of language learners in their field of study or to train them in discourse strategies specific for a particular register.*

Keywords: *Maritime English; spoken corpus; written corpus; subvariety; corpus-based analysis;*

1. Introduction

Registers and genres have been widely and extensively studied from various perspectives. Corpus-based research has provided empirical basis for a descriptive study of register and genre, touching upon various features, from describing stance, hedging devices, lexical and grammatical characteristics, to more comprehensive multidimensional studies conducted by Biber. The studies have shown that language characteristics differ dramatically from one register or genre to the next, indicating a specific relation between language units and their pragmatic or discourse functions. This had wide implications for education and teaching the specific kinds of language that learners will need. Consequently, the linguistic characteristics of these specialized registers had to be fully described before developing suitable teaching materials and methods. Furthermore, the users need to be aware of the specific features of a register or genre in order to use them in specific contexts or to make alterations within them to achieve a specific effect.

The goal of this study is to demonstrate how linguistic variation exists across subvarieties of Maritime English based on corpus evidence. Such a holistic overview and comparison of Maritime English subvarieties has not been conducted in this way, so this study is a contribution to gaining a wider understanding of the domain. The first hypothesis was that linguistic structures will be used in each subvariety in line with the communicative purpose of the particular register (e.g. clarity and efficiency in maritime communications or precision and formality in maritime law). The second hypothesis was that the spoken Maritime English will show different patterns from written subvarieties, particularly in syntax and modality. The third hypothesis was that each subvariety will have some domain-specific features, regarding the frequency of words, semantic classes of verbs and nouns, use of pronouns, etc.

2. Literature review

2.1. Register and genre

In literature, both register and genre have been used to refer to varieties associated with particular situations of use and particular communicative purposes. In most studies one of these terms is adopted, and the other one disregarded, or no distinction is made between the two. However, some scholars do make a fine distinction by using the term register to refer to a general kind of language associated with a domain of use, such as a legal register, scientific register, while genre refers to a culturally recognized message type with a conventional internal structure, such as a scientific article, a business memo, or a sales contract. Following that, the research on register focuses on lexico-grammatical features and genre focuses on socio-cultural actions, and refers to "how things get done, when language is used to accomplish them" (Martin 250). Biber, who has conducted extensive research on the subject, used the term 'genre' in his early works, and later on used the term 'register'.

This article will not deal with theoretical considerations regarding these two terms, but it was necessary to make this distinction based on past research to define the focus of the study. Here, the term 'register' was adopted as the focus is on the linguistic features and their variation depending on the context of the situation, relationship between speakers, and not on text structure or organization even though the purpose of communication, generally associated with genre, was mentioned in discussing the findings.

2.2. Research of the Maritime English register

In the area of Maritime English, several subvarieties are distinguished such as Maritime Legal English, Maritime Business English, Marine Engineering English. Some corpus-based studies have been conducted aiming to describe

these specific subvarieties. Pritchard (2007) explored the concept of minimum vocabulary requirements in Maritime English for various educational and professional levels, including BSc courses and STCW competency standards. The study distinguished different types of vocabulary such as core, technical, semi-technical, and general English, and emphasized the importance of structuring vocabulary learning by frequency and relevance to maritime contexts. The research involved a corpus-based quantitative analysis of three maritime text registers – navigation, maritime law, and marine engineering – showing the distribution of vocabulary across general service lists, academic word lists, technical words, and low-frequency maritime-specific words. Findings suggest that roughly 50% of vocabulary in maritime texts consists of general English words, with a significant portion being technical or maritime-specific, highlighting the need for targeted vocabulary teaching. The paper concludes by advocating for the creation of a comprehensive maritime English corpus and further research to define vocabulary levels, compounds, collocations, and lexical sets for effective learning and safety at sea.

Further studies have also been conducted in the specific subvariety of Maritime VHF Communications at the macro-perspective of its structure and at a discourse level. The studies mostly focused on discrepancies between the standard protocol of communication and real communication exchanges, (e.g. Jurkovič). John, Brooks and Schriever (2017) also quantitatively profiled the speech patterns of bridge team communication by non-native English speakers in maritime settings and compared these patterns to non-nautical communication. The study analyzed vocabulary growth, word frequencies, lexical and key word densities, and grammatical diversity using transcripts from full-mission ship handling simulations against established corpora like the Brown Corpus, the Vienna-Oxford International Corpus of English, and the Standard Marine Communication Phrases (SMCP). The study found that bridge team communication has a significantly restricted vocabulary size and distinctive word frequency distributions compared to general English communication, but it closely aligns with the prescribed SMCP language. Lexical density in bridge team communication is higher than in non-nautical speech by non-native speakers and features a higher proportion of nautical key words, indicating its idiomaticity within the maritime domain. Ultimately, the research provides a quantitative linguistic profile that models the unique discourse of bridge team communication, offering a tool to assess and improve maritime English training and safety.

Kegalj (2024) analyzed a corpus of 30 routine maritime VHF communication exchanges from the Adriatic area to identify which message elements and syntactic structures are most frequently omitted. The study focused on pragmatic omissions that occur due to language economy, despite guidelines advocating full, clear communication. The analysis revealed that

omissions are common, particularly of message parts like "address" and "This is," and parts of speech such as personal pronouns, auxiliary verbs, and articles. Ships more often omitted subjects, personal pronouns, and auxiliary verbs, while shore stations omitted more predicates, full verbs, and question words. Importantly, these omissions did not hinder understanding due to shared contextual knowledge and cooperative communication, supporting safe and effective communication through predictable scripts.

Se-Eun Jhang and Sung-Min Lee (2013) conducted a study on clusters and key clusters in a Maritime English corpus. The study aimed to identify the number of 4-gram clusters in the Maritime English Corpus (MECO II) and determine how many of these are associated with general English by comparing them with the British National Corpus (BNC) Baby. It also sought to explore the statistical relationship between the syntactic categories and semantic functions of key clusters using chi-square tests and to analyze their semantic functions such as stance bundles, discourse organizers, referential expressions, and special bundles. The findings revealed that 78% of the clusters in MECO II are specific to Maritime English, with 22% overlapping with general English, and that special bundles are a unique functional characteristic of Maritime English not found in general English. Chi-square tests showed significant relationships between syntactic structures and semantic functions of key clusters. Additionally, notable differences in collocational patterns and dispersion plots between Maritime English and general English clusters underscored the specialized nature of Maritime English expressions.

Franceschi (2014) illustrated the linguistic features of Maritime English across different contexts, highlighting its role as both a specialized academic/professional discourse and a simplified vehicular language for communication at sea. The study analyzed written texts from marine engineering, marine electronics, and maritime law, as well as spoken data including a ship-to-shore conversation and the IMO's Standard Maritime Communication Phrases. The findings revealed that written Maritime English is structurally complex, especially maritime law texts, with a mixture of semi-technical and highly specialized terminology, formal style, and some archaic expressions. Spoken Maritime English, by contrast, is characterized by simplification, standardization, and a focus on clarity and immediacy, often reflecting features of English as a lingua franca. Overall, the findings of the study show that Maritime English is a multifaceted language variety serving different communicative purposes with distinct lexical, syntactic, and pragmatic properties.

Wenyu Lu, Sung-Min Lee and Se-Eun Jhang (2017) aimed to identify typical linguistic features that distinguish international public maritime institutional law texts from private maritime institutional law texts using

corpus-based methods. By constructing two specialized corpora and comparing them with a general English reference corpus, the study analyzed key words, key clusters, and key semantic domains to reveal distinctive patterns. Key findings include the complementary distribution of terms such as *regulation/administration* in public law versus *liability/compensation* in private law, and the dominance of modal verbs *shall* and *may* in both corpora. Additionally, specific four-word key clusters and semantic fields were found to be uniquely associated with public law, while other clusters appear only in private law. These linguistic distinctions offer a practical means for non-legal practitioners to differentiate between public and private maritime legal texts and potentially other legal documents.

3. Methodology

This corpus-based study focuses on identifying salient linguistic patterns in a register and then comparing four registers according to the linguistic features defined by the patterns. Four corpora were compiled from different fields of Maritime English: maritime law, marine engineering, maritime communications and academic maritime English. After compiling the corpora, corpus data on word frequencies, parts of speech frequencies, noun, verb and pronoun frequencies, keyword and key term lists, and n-grams were extracted using standard tools to identify the recurring features or patterns. The corpora were also analyzed based on quantitative measures such as type-token ratio (TTR), standardized type-token ratio (STTR), average sentence length (AVS), noun/verb ratio (N/V), and lexical density calculated as a ratio of lexical words versus total number of tokens (Stubbs 71-73) and as a ratio of lexical and functional words (Halliday 61-64). These data indicate lexical diversity of the studied corpora, syntactic complexity and informational density, while the data about word distribution within the corpora indicate some potential features of the four registers. In addition, the corpora were also compared in terms of their level of readability to establish the level of their complexity. The difference in corpus size was mitigated through normalization. Finally, the 100 most frequent lexical items and n-grams were categorized semantically or functionally, and results were triangulated through qualitative analysis.

After quantitative analysis, the results were qualitatively analyzed and categorized. The one hundred most frequent nouns, verbs, keywords and key terms were categorized according to semantic categories, while verbs were also categorized according to grammatical category of tense, and n-grams were categorized according to their function. The results were triangulated through qualitative analysis. Finally, the corpora were compared and their features described.

3.1. Corpora

Maritime English is a subdomain of occupational English with several subvarieties: English for navigation and maritime communications, English for maritime commerce, English for maritime law, English for marine engineering and English for shipbuilding (Bocanegra-Valle 3570). This study included one spoken and three corpora of written texts from the domain of Maritime English, that represent different registers and therefore, it is expected that their linguistic features will differ to a certain extent.

MARCOM, a spoken corpus of maritime routine communication between ships and shore stations, includes speakers producing unconstrained messages in authentic routine situations. About ten hours of authentic recordings were granted by the Croatian Maritime Safety Directorate of the Ministry of the Sea, Transport and Infrastructure, provided that the GDPR requirements are met. The corpus contains 93920 tokens, which roughly amounts to 500 VHF communications between different ships and VTS stations. This involved protecting the original recordings and anonymizing the transcripts by removing all names, geographical locations and any other information that could lead to identification. This register is quite specific as it is constrained on the one hand with real-time production circumstances and on the other by various documents issued by international organizations such as International Maritime Organization (IMO), the International Association of Lighthouse Authorities (IALA) and the International Telecommunication Union (ITU), which regulate its use, the length of utterances, terminology used and recommend specific language structures. The participants communicate exclusively over the radio, so paralinguistic features, such as body language and facial expressions, do not play a role. Participants rely on intonation, closed communication loop and contextualization and knowledge of the world to achieve understanding. This register shares some features of everyday spoken discourse, such as simpler syntax, spontaneity and contextualization (cf. Schober and Brennan 125-128, Cornish 227-230) and features of specialized discourse, such as specialized and restricted vocabulary, formal register, standardized phrases, repetitions and confirmations (cf. Franceschi 82-84).

MARLAW is a corpus of legal texts covering legal aspects of shipping, ships' contracts, limitation of liability, insurance claims, ships' ownership and registration, various maritime events, safety and marine pollution control. This corpus contains 564267 tokens, from 14 international conventions and five international regulations, excluding EU maritime regulations. The main reason for this was that research have shown that EU English differs from British legal English (cf. Trosborg 1997, Koskinen 2000, Tosi 2005, Biel 2014) and some scholars consider it as a different variety of legal English.

MARENG is a corpus of written technical texts that deal with the design, operation, construction and maintenance of ship engines and machinery or guidelines for repairs. It is the largest of the four corpora and contains 2,687,102 tokens.

The corpus MARAC contains scientific articles from international maritime conferences IAMU. It has 670805 tokens and represents the academic register dealing with topics from the area of maritime sector from a scientific perspective.

The corpora all belong to the domain of Maritime English, but different registers, utilizing different linguistic structures to express ideas and following different linguistic norms. In that sense, it is expected that the results of the corpus study will pinpoint the specific features of each register and indicate similarities and differences among the registers within one domain.

3.2. Description of data analysis

The paper describes the investigation of patterns of linguistic variation among different registers within one domain. The study included the analysis of word frequencies to see the differences between high-frequency and low-frequency, i.e. specialized words for each register. This gave an insight into the vocabulary patterns across registers and the relation between word use and a particular register, which might have implications for language learning. The analysis further focused on parts of speech frequencies as preferences in that sense may reveal tendencies within the register. For example, a higher frequency of nouns and prepositions indicates greater tendency towards formality, while informal texts show higher use of pronouns and verbs.

In addition, nouns were further categorized into semantic categories, according to whether they referred to things, persons, places, quantity, concepts or actions, which indicated the tendencies within certain registers. The same was done for verbs, which were categorized into existence verbs, mental verbs, activity verbs, communication, aspect, occurrence and causative verbs. This semantic analysis is useful to gain insight into the dominant categories and central themes within the register. The verbs were also analyzed according to tense and aspect to gain insight into the distribution of these categories within registers as various registers tend to prefer different tenses, e.g. spoken conversation shows a more varied use of tenses, while academic texts prefer present simple and passive.

These tendencies were further investigated in the analysis of n-grams, i.e. frequently co-occurring sequences of words. The data about n-grams reveal patterns and identify recurrent phrases typical for a register. Considering the size of the corpora, the frequency cut-off point was set at five, as the corpus MARCOM is rather small and would probably not demonstrate

significant n-grams at a higher cut-off. The n-grams were categorized according to their discourse function into stance, expressing attitude, referential, i.e. those that make a reference to some entity, and discourse organizing n-grams, i.e. those that express relations between parts of sentences (cf. Biber 138-148). The data on keywords and key terms was classified into semantic categories thus providing insight into register-specific vocabulary and topics.

The quantitative data provided information on lexical density and complexity of individual corpora providing additional insight into the features of the corpora, like lexical richness, variety and repetition within a register. Type-token ratio (TTR) and standardized type-token ratio (STTR) were both included as TTR tends to be influenced by the size of the corpus, so STTR was used to compensate for size. The data was complemented with the Flesch-Kinkaid reading ease as a measure frequently used in assessing the complexity of a text, giving a score between 1 and 100, with 100 being the highest readability score. These data allow us to compare registers based on their lexical diversity.

Based on all of the mentioned data together, the features of each register can be described and the registers can be compared in terms of their formality, tendency towards nominalization, complexity, information density, etc.

4. Results

The data in Table 1 shows the quantitative indicators of lexical diversity. According to that, the MARCOM corpus shows the lowest TTR and STTR values. The MARAC corpus shows the highest value, followed by MARENG and MARLAW corpora. The MARLAW corpus showed the greatest value in average sentence length (AVS), while the lowest AVS value was calculated for the MARCOM corpus, which is expected as this is a spoken corpus. The other two lexical diversity scores also yielded interesting findings. Halliday's ratio of content words and total number of sentences is the highest in the MARAC corpus, while the MARCOM corpus showed the lowest value of this ratio. Stubbs' ratio of content words and total number of tokens shows similar results in the three written corpora, while the MARCOM corpus again stands out with the lowest score here. The ratio of content versus functional words shows that MARLAW and MARCOM use more functional words than the other two corpora, which are similar in that sense. Finally, the ratio of nouns and verbs shows the greatest values in the MARAC corpus, followed by MARCOM and MARENG, while MARLAW shows the lowest value of the ratio.

	MARENG	MARCOM	MARLAW	MARAC
TTR	75	65.6	82	78.4
$STTR_{100}$	66	55	61	70
AVS	19.17	11	61.5	23.32
N/V	2.99	2.98	2.59	3.23
CONT/FUNC	1.14	2.1	2.54	1.13
LD Hal	13.59	7.02	11.47	15.79
LD Stubbs	53.3	41.7	53	53.1

Table 1. Quantitative data about the corpora (source: author).

Table 2 shows the readability scores for the four corpora. The most popular score, used even in Microsoft Office Word documents, Flesch-Kincaid reading ease indicates that MARCOM is fairly easy to read, while MARENG with a score <30 is at a college graduate level, best understood by university graduates and very difficult to read. MARAC and MARLAW fall within the college level range and are also difficult to read. The Gunning Fox Index indicates the years of formal education a person needs to understand a text on the first reading. This corroborates the results of the previous score. The SMOG index is a popular formula indicating the difficulty of a text based on the number of words with three or more syllables in a sample. It still shows that MARCOM is the simplest, but gives no preference to the other three corpora, which might indicate the prevalent use of polysyllabic words in those corpora.

READABILITY SCORES	MARENG	MARCOM	MARLAW	MARAC
Flesch-Kincaid Reading Ease	24.9	78.4	40.2	47.1
Gunning Fog Index	18.8	4.6	16.3	11.7
SMOG Index Score	12	6.8	12	12

Table 2. Readability scores for the corpora (source: author).

The analysis of parts of speech (POS) at three frequency levels, shown in Figure 1, indicates that the majority of different words (1 – 20 hits in the corpus) are used in the MARENG corpus, while MARCOM and MARLAW show the least number of different words. Moderately common word types show a similar pattern in the corpora.

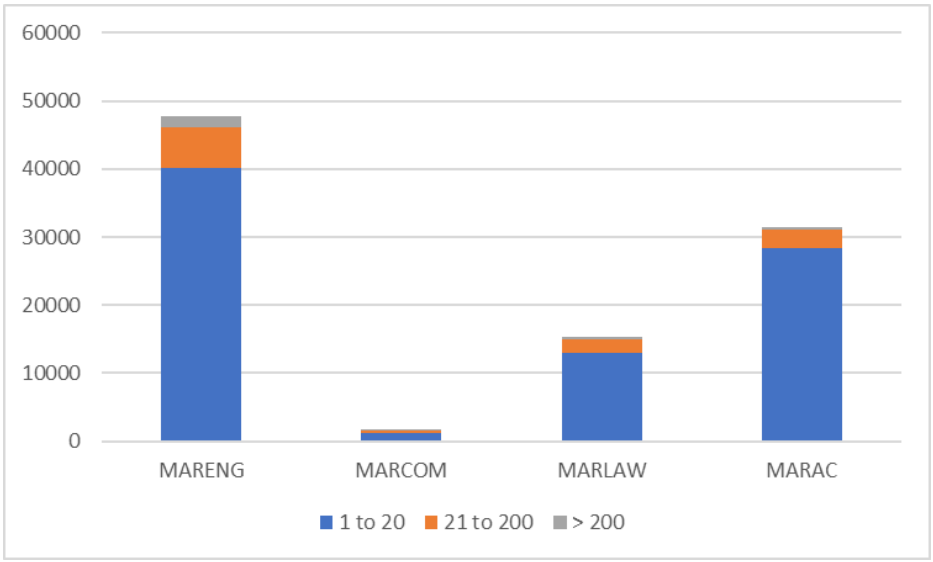

Figure 1. Analysis of POS at three frequency levels

The analysis further focused on individual POS frequencies, as shown in Figure 2, normalized per million words. The corpus results have shown that nouns prevail in all corpora, with MARAC and MARLAW having the greatest share of nouns. The MARENG and MARLAW corpora show a similar distribution of verbs, adjectives, determiners and prepositions, with MARAC following this tendency. MARCOM interestingly shows the greatest share of pronouns and numbers and a smaller share of conjunctions, determiners and prepositions.

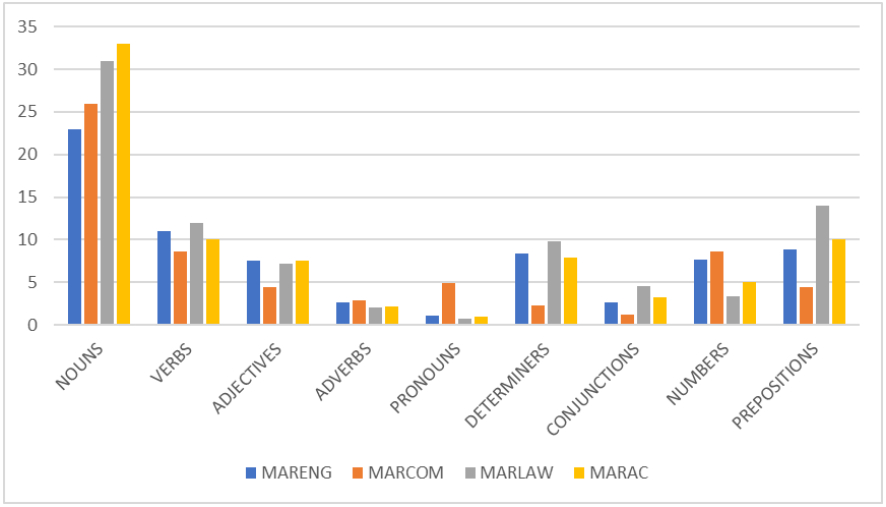

Figure 2. Analysis of POS frequencies in corpora.

As noun phrases contain much of the referential information, nouns were further analyzed into following semantic categories: things (e.g. engine, valve, ship), persons (e.g. seafarer, student, officer), places (e.g. port, city, area), quantity (e.g. volume, year, ratio), concept (e.g. safety, system, information) and action (e.g. transport, learning, operation), as shown in Figure 3. The results show that the MARENG corpus focuses more on nouns indicating things, places and concepts, without reference to persons, the MARCOM corpus focuses more on persons, places and quantity, the MARLAW corpus focuses on things, persons, concepts and actions, while the MARAC corpus focuses on persons, concepts and actions, almost no reference to places.

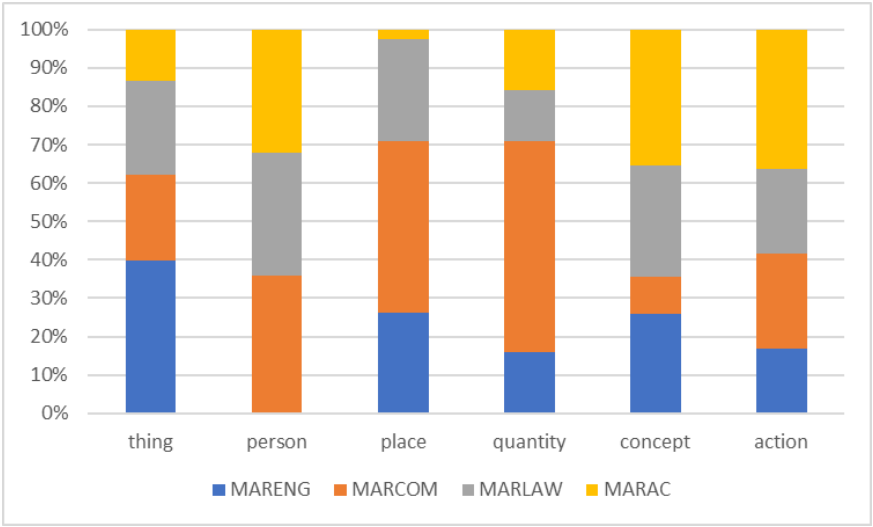

Figure 3. Semantic analysis of nouns in the corpora.

As the verbs were the second most frequent POS, they were also grouped into major semantic classes, as shown in Figure 4. The semantic classes included the following: existence (e.g. *be, become*), mental (e.g. *identify, relate*), activity (e.g. *enter, come, berth*), communication (e.g. *call, tell, repeat*), aspect (e.g. *have, keep*), occurrence (e.g. *depend, check*) and causative (e.g. *affect, ensure*).

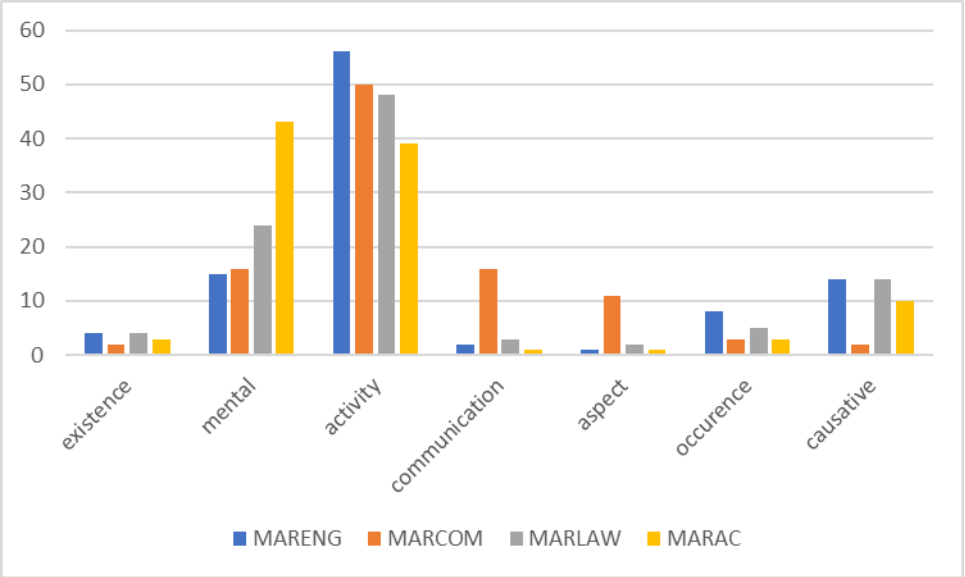

Figure 4. Semantic classes of verbs in the corpora.

Overall, mental and activity verbs are the most frequent in the corpora. As expected, the MARCOM corpus relies more on communication and aspect verbs than other corpora, with very few causative verbs. MARENG shows a heavy reliance on activity, occurrence and causative verbs. In MARLAW, the most frequent verbs are mental, activity and causative, while MARAC relies heavily on mental verbs, followed by activity and causative verbs.

Verbs were also considered in relation to the variation in the marking of tense, aspect, and modality (Figure 5). Tense distribution might provide information on syntactic sophistication. All corpora showed a preference towards the present tense and the simple aspect, MARENG in particular, expressing change and gradual development and actions happening at the time of speaking (e.g. *Designers are introducing technical and system solutions…*), or facts (e.g. *The improvements lead to a finer approach to ship stability*) within this informational register. Among the marked tenses, MARCOM used the progressive aspect for present activities (e.g. *I'm entering sector 4*) and the future (e.g. *Then I will proceed a bit more South*) more than the other corpora. Another distinctive feature is the use of modal verbs (particularly ‘shall’, e.g. *Parties shall endeavour to co-operate*) in MARLAW which is far greater than in the other corpora.

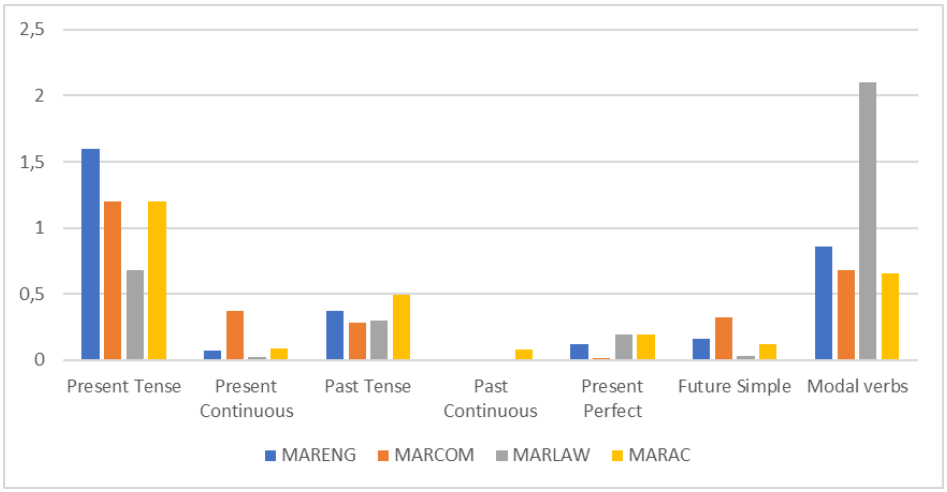

Figure 5. Distribution of tenses in the corpora.

Another marked feature analyzed in the study was the use of pronouns. Table 3 shows a list of pronouns used in the corpora given according to their frequency in the corpora. The MARCOM corpus stands out in this sense, with a prominent use of second person pronouns, followed by first person pronouns. MARLAW, MARENG and MARAC show a more prominent tendency towards the use of more neutral third person pronouns.

MARENG	MARCOM	MARLAW	MARAC
it	you	it	it
you	your	their	their
its	we	its	they
they	I	they	we
your	me	them	its
we	my	her	them
their	it	I	our
I	our	his	you
our	us	itself	I
them	they	she	his

Table 3. List of pronouns according to frequency in the corpora.

A further analysis was performed on keywords and key terms, extracted in Sketch Engine, based on their keyness score. Keyness score compares the frequencies of words, or single-token items, and multi-word expressions in the focus corpus with the frequencies in the reference corpus (in Sketch Engine the enTenTen21 corpus is set as default) and thus identifies the words that are typical, i.e. more frequent in the focus corpus than in the reference corpus. Keywords and key terms are useful for identifying the topic of the corpus. The first 100 most frequent keywords and key terms in each corpus were classified according to the domain they refer to. As these are nouns or noun phrases, a similar categorization was used as previously described for nouns. The results are shown in Figure 6.

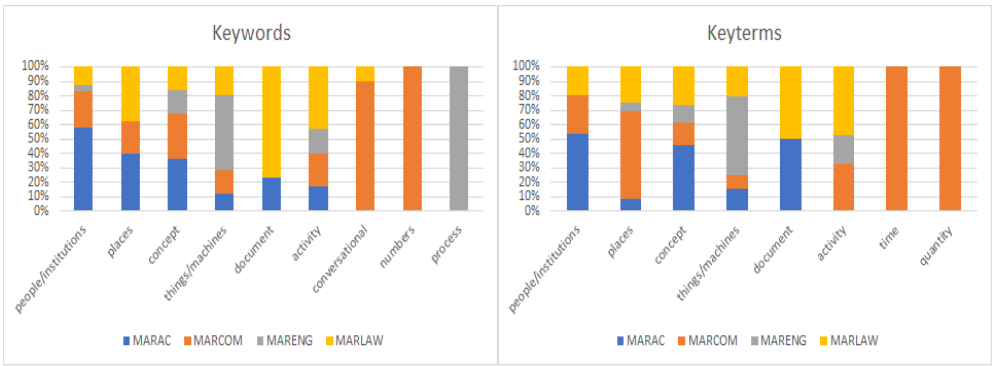

Figure 6. Results for the most frequent keywords and key terms in the corpora.

The results indicate that MARAC register frequently refers to people or institutions, concepts and documents. The distinguishing vocabulary items in MARCOM refer to places, numbers, time, quantity or people. There are also frequent keywords that perform some pragmatic conversational functions, e.g. greetings, wishes, etc. In MARENG the key vocabulary refers to things and machines (e.g. *piston, crankshaft, camshaft*), or processes (e.g. *combustion, scavenging*), while MARLAW key vocabulary indicates documents (e.g. *STCW, annex*) and activities (e.g. *watchkeeping, fire-fighting*).

The final part of the analysis included the analysis of n-grams which were extracted in Sketch Engine and then analyzed according to their discourse function (cf. Biber 139) into stance, referential, discourse and conversational n-grams (Figure 7). In this respect, MARCOM has shown specific features as it contains the most referential n-grams and n-grams performing some conversational function. MARAC and MARLAW have shown similar features, while MARENG differs in terms of referential n-grams.

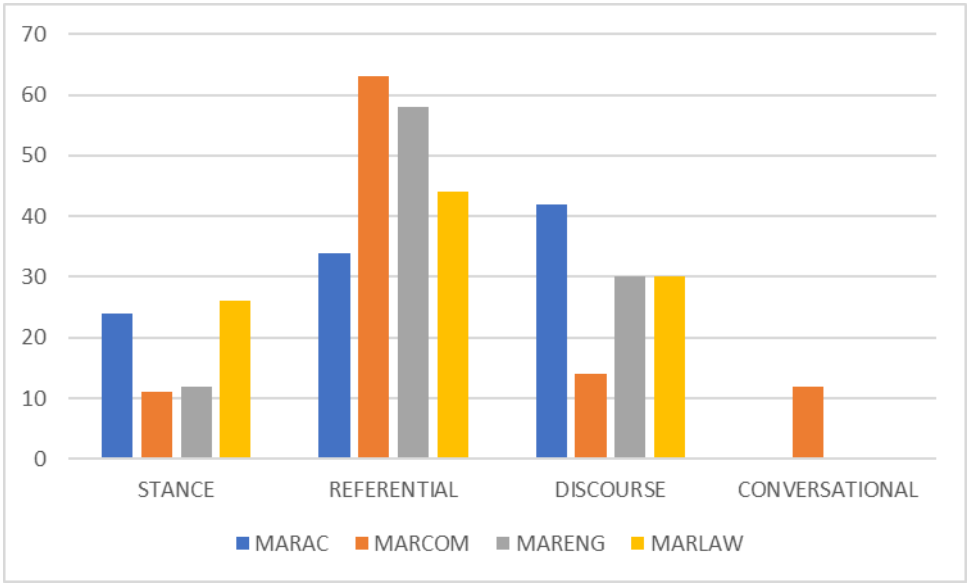

Figure 7. Distribution of n-grams in discourse categories.

5. Discussion

The quantitative data about lexical diversity and readability across the registers show interesting findings. According to that, MARAC corpus has the highest STTR score and consequently a greater complexity as it has more lexical words per sentence. Halliday's score of lexical diversity also indicates more complexity of this corpus than others, which might imply that less functional words are used in sentences than in the other corpora. It also has a greater tendency towards nominalization and high informativity than the other corpora, confirmed by the noun and verb ratio. This corpus is also quite complex when it comes to average sentence length and the Stubbs' score of lexical density. In terms of readability, it falls within the college level range, i.e. it is rather difficult to read, requiring at least 12 years of formal education to understand such a text. It also has a rather high share of relatively infrequent words, indicating a variety in lexicon and diversity in used word forms. The distribution of POS shows the prevalence of nouns in the corpus, with nouns, keywords and key terms referring to persons, concepts and action being most frequent. This is in line with the communicative function of the corpus, focusing on these particular topics. Similarly, the distribution of semantic verb classes aligns with the communicative function, relying on verbs of thinking, action and cause in the present or past tense. This indicated cognitive processes, present or past, and present actions and cause-effect relations between concepts. This is also in line with n-grams, expressing stance,

reference or having a discourse function, as these texts frequently express opinions, refer to other concepts and make references among concepts.

The corpus MARLAW, as expected, has the highest score when it comes to sentence length and a rather high value of lexical density indicating complexity of the corpus. This is further corroborated by the readability scores, as well as the other lexical density measures. What further distinguishes this corpus from other complex corpora is the greater use of functional words, which indicated a greater level of grammatical cohesion within the texts. This might be the result of the fact that these texts usually have a lot of intertextual and intratextual referencing, as well as the need for explicit expression of relation among concepts. This is also corroborated by the data on n-grams, where referential and discourse n-grams are most frequent in the corpus. MARLAW is also specific for the distribution of POS at three frequency levels, which showed that it uses fewer number of different words. In legal English it is important to use the same expressions for the same contexts, without variations in terminology to reduce ambiguity or misunderstandings. The corpus shows a greater tendency towards nominalization and consequently might imply greater compaction of information. The nouns and keywords in the corpus refer to things, persons, concepts and actions, and key terms also refer to documents, which are frequent topics in legal discourse. The analysis of verbs indicates that modal verbs are rather frequent, reflecting the normative nature of legal texts. This modality probably originates from the nature of the text providing obligations, permissions, bans. Furthermore, the data on pronoun frequency indicates an indirect nature of texts within the corpus, distancing the sender of the message from the receiver.

The data about the MARENG corpus show some distinguishing features of Marine Engineering texts. The high ratio of hapax legomena indicates a greater reliance on rare, highly specialized words in the corpus. Together with the data on readability, which places this corpus into the category of texts that are very difficult to read, suggests that the texts in the MARENG corpus are complex with highly dense content. This is further corroborated by the data on noun-verb ratio, showing the tendency of the corpus towards nominalization and informativity. The distribution of parts of speech categories shows that the corpus has a high frequency of content words, but also function words, such as determiners, conjunctions and prepositions, which suggests a more complex syntax in the texts. This may be related with the data on n-grams, where the referential n-grams were the most frequent in the corpus. These data indicate a high cohesion and intratextual referencing in the corpus. The semantic analysis of nouns and keywords shows that the topic in the corpus focus on things and concepts, which is in line with the communicative function of the corpus. The semantic analysis of verbs in

MARENG shows a heavy reliance on activity, occurrence and causative verbs, also in line with the communicative function of the texts describing events and activities, and cause-result relations. Similarly to MARLAW and MARAC, it most frequently uses the third person pronoun to indicate indirectness, but the second person pronoun is also very frequent, probably in the context of instructions or guidelines which take a more direct approach towards the receiver of the message. When it comes to verb tense, the present tense is the most frequent one in this corpus, probably due to the fact that the texts express facts and general information.

Finally, the MARCOM corpus, as the only spoken corpus in this study, shows some features that distinguish it not only from the other corpora in the study, but also from spoken corpora in general, proving its specialized nature. In terms of complexity, MARCOM is the least complex, with the lowest average sentence length and reduced lexical density measures, indicating simpler texts in comparison to other corpora. This also indicates a greater repetition of words, which is in line with the communicative purpose of the corpus – simplicity, clarity, reduction of ambiguity. This is further corroborated by the data on hapax legomena, that indicates a lower word variety and limited vocabulary in line with the standardization and institutionalization of maritime communications. When it comes to POS distribution, MARCOM spoken corpus is rather specific as spoken corpora usually show either the prevalence of verbs or an equal share of nouns and verbs, but MARCOM shows a high noun-verb ratio. According to Biber (2006), nouns and verbs are used approximately to the same extent in spoken registers, and adjectives and adverbs are distributed in a similar way, i.e. nouns and adjectives are more common in written registers, and verbs and adverbs are more common in spoken registers. In case of MARCOM, the situation is somewhat different, mostly because MARCOM is not a typical spoken corpus. This feature distinguishes this corpus from general language spoken corpora, indicating its high informativity and tendency towards nominalization. In line with the communicative function of the corpus, it also shows a large share of numbers, which appear in the context of providing information on ship position, time of arrival or departure, quantity of cargo or fuel on board, etc. It demonstrated frequent use of function words indicating a certain degree of simplification, but also rather explicit grammatical relations. Another distinguishing feature of this corpus is the high share of present progressive, future tense and modal verbs in relation to other corpora in the study, indicating its focus on instantaneous activities, actions at the time of speaking or future actions, as well as obligations and permissions expressed by the authorities when a ship needs to do something or is granted permission to anchor, berth or enter the port. A further specific feature of the corpus is the high use of pronouns, most frequently first and second person pronouns, in

comparison with other corpora, indicating directness and interactivity of the register, second person pronouns being used specifically for directive purposes. MARCOM interestingly shows the greatest share of numbers and a smaller share of conjunctions, determiners and prepositions, which is in line with the communicative function of the corpus. The semantic analysis of nouns in MARCOM showed the dominance of the categories of persons, places and quantity, while the semantic analysis of verbs showed the prevalence of communication and aspect verbs, which is all in line with the communicative function of the corpus, focusing on ports of call, ports of destination, berths, anchorages, and communicating about those. These findings can be linked with the data on n-grams, in which referential and conversational n-grams are most frequent in the corpus, corresponding to its communicative function.

6. Conclusions

Based on quantitative data from the corpora, the findings of the study demonstrate how the examined subvarieties use linguistic means to achieve different functions. The MARCOM corpus has shown greater distinction from the other corpora, which primarily stems from the fact that it is the only spoken corpus, so it utilizes different linguistic strategies than written corpora to perform its communicative function. The study has also shown that it differs from spoken genres in other domains, with its specific traits like the frequency of nouns and verbs, frequency of numbers, frequency of specific verb and noun semantic categories. As expected, it showed the least lexical diversity and the highest level of readability in all scores. The frequency results also showed a greater share of pronouns, indicating the directive and interactive nature of the register. The findings correspond to the communicative function of the corpus which involves short and clear messages about current events, providing information to ships or shore stations about cargo, time of arrival or departure, reporting the position or condition, etc. A specific trait of the register is the pragmatic conversational elements visible in key terms and n-grams as the most frequent and the most distinguishing linguistic units in the register. Even though the register is supposed to convey strictly formal relations, it still draws on the resources from everyday spoken language to perform social interaction.

The other three corpora showed both similarities and differences. When it comes to lexical density and complexity, academic discourse, which brings new insights and research results, uses many content words so it is expected that it will be highly informative and rather complex. Legal language usually involves sentences with many insertions and additions and are thus rather long. Engineering is a specialized domain, so the texts can be

understood by specialists in this area, with years of education, training and experience. All three written corpora showed a pronounced tendency to nominalization and a frequent use of polysyllabic words, which is why their readability scores fell into the category of "difficult to read" or "very difficult to read". Each register focuses on a specific semantic domain, characteristic of that register, related to their communicative function. In terms of tense and aspect, all corpora demonstrate their immediacy, ongoing relevance of information and direct nature of discourse, e.g., MARLAW demonstrates a specific feature regarding the pronounced use of modal verbs. In terms of text cohesion, demonstrated by n-grams, MARLAW and MARENG had a similar distribution of n-grams, with a larger share of stance n-grams in comparison with the other two corpora. This might be attributed to the fact that expressing stance is characteristic for the communicative function in these two domains. As opposed to that, MARENG has quite a low share of stance n-grams, as it focuses more on referential and discourse n-grams referring to specific things and objects and expressing relations between them. MARCOM contains mostly referential and conversational n-grams which corresponds to the features of the register mentioned above.

In conclusion, the corpus-based analysis demonstrated that subvarieties of Maritime English utilize linguistic structures in a specific way to achieve their communicative function. Therefore, the first hypothesis was confirmed. The second hypothesis about the MARCOM corpus was also confirmed, with another relevant finding about this corpus being different from general spoken corpora as well. The third hypothesis about domain-specific features was partially confirmed, as the corpora showed both similarities and differences in different areas under study.

This study has demonstrated how corpus-based data can serve in describing the features of different registers within one thematic domain. Each corpus exhibited some of the features typical of that register, but also features that are quite specific owing to the specialized nature of each corpus. Such insights might be useful in teaching Maritime English as they demonstrate the actual communicative needs and provide real-life examples of language use based on corpus data. Furthermore, teaching functional and pragmatic aspects of language may help students understand not just the content, but also the ways and the reasons why some structures are used in different contexts.

Works Cited

Biber, Douglas. *University Language: A corpus-based study of spoken and written registers*. Amsterdam/Philadelphia: John Benjamins, 2006.

Biel, Łucia. "The textual fit of translated EU law: a corpus-based study of deontic modality." *The Translator 20/3* (2014): 332-355.

Bocanegra-Valle, Anna. "Maritime English." *The Encyclopedia of Applied Linguistics.* Ed. C.A, Chapelle. Oxford: Wiley-Blackwell, 2012. 3570-3583.

Cornish, Francis. "Understanding spoken discourse." *Encyclopedia of Language and Linguistics*. 2nd ed. vol 13. Elsevier, 2014. 227-230.

Franceschi, Daniele. "The Features of Maritime English Discourse." *International Journal of English Linguistics* 4/2 (2014): 78–87.

Halliday, Michael A. *Spoken and written language.* Geelong Vict.: Deakin University, 1985.

Jhang, Se-Eun, Sung-Min Lee. "Clusters and key clusters in the Maritime English Corpus." *Journal of Language Sciences* 20/4 (2013): 199-219.

Martin, James R. "Process and text: two aspects of human semiosis." *Systemic Perspectives on Discourse, Vol. 1*. Eds. James D. Benson and William S. Greaves. Norwood, NJ: Ablex, 1985. 248-274.

John, Peter, Brooks, Benjamin, Ulf Schriever. "Profiling maritime communication by non-native speakers: A quantitative comparison between the baseline and standard marine communication phraseology." *English for Specific Purposes, Volume 47* (2017): 1-14.

Kegalj, Jana. "Deliberate pragmatic omissions in maritime VHF communications." *Proceedings of the 29th International Maritime Lecturers Association (IMLA) Conference* (2024): 483-489.

Koskinen, Kaisa. "Institutional Illusions. Translating in the EU Commission." *The Translator 6(1)* (2000): 49-65.

Lu, Wenyu, Lee, Sung-Min, Se-Eun Jhang. "Keyness in maritime institutional law texts." *Linguistic Research 34(1)* (2017): 51-76.

Pritchard, Boris. "Minimum (technical) vocabulary – some issues in Maritime English." *Proceedings of the19th International Maritime English Conference (IMEC 19) "The Human Element in Maritime Accidents and Disasters - a Matter of Communication"*. Rotterdam: Boston (MA): Taipei: Rotterdam Maritime College, 2007. 70-86.

Schober, Michael F., Susan E. Brennan. "Processes of Interactive Spoken Discourse: the Role of the Partner." *Handbook of Discourse Processes.* Eds. Arthur Graesser, Morton Ann Gernsbacher, Susan R. Goldman. New Jersey: Lawrence Erlbaum Associates, 2003. 123-164.

Stubbs, Michael. *Text and corpus analysis: Computer assisted studies of language and culture.* Oxford: Blackwell, 1996.

Tosi, Arturo. „EU Translation Problems and the Danger of Linguistic Devaluation." *International Journal of Applied Linguistics 15(3)* (2005): 384-388.

Trosborg, Anna. *Rhetorical Strategies in Legal Language: Discourse Analysis of Statutes and Contracts*. Tubingen: Gunter Narr Verlag, 1997.