30th
International Maritime Lecturers Association Conference
Split Croatia
www.imla30.org

International Maritime Lecturers Association
FACULTY OF MARITIME STUDIES

ASSESSING THE RELEVANCE OF ZIPF'S LAW IN MARITIME TECHNICAL TEXTS

Zorica Đurović*

Abstract. In the early 20th century, the American linguist George Kingsley Zipf identified a linguistic regularity in the relationship between word frequency and rank, later known as Zipf's Law. This principle suggests that the product of a word's frequency and its rank remains approximately constant. While numerous studies have confirmed its validity across various natural languages and domains, its applicability to specialized technical texts remains an area of interest. This study examines the vocabulary frequency and distribution in marine engineering instruction books to assess the extent to which Zipf's Law holds in this highly specialized linguistic domain. Using a corpus of marine engineering texts, we analyzed the most frequently occurring words and compared the findings with existing research on different text types. The results provide insights into the approximate applicability of Zipf's Law to technical corpora and highlight specific linguistic characteristics of such texts. Additionally, this study underscores the need for a multidisciplinary approach and collaboration in corpus linguistics research. Moreover, it lays the groundwork for a broader and more comprehensive investigation of maritime language corpora.

Keywords: Vocabulary; Frequency; Zipf's Law; Marine Engineering, Instruction Books.

University of Montenegro,
Faculty of Maritime Studies
Kotor, Put I Bokeljske brigade 44, Montenegro

* **Corresponding author:**
zoricag@ucg.ac.me

https://doi.org/10.7225/imla30.50

1. INTRODUCTION

Postulating that (natural) language, similar to other natural phenomena, tends toward a state of equilibrium, George Kingsley Zipf formulated the "Principle of Least Effort" to account for the frequency distribution of linguistic elements. This theoretical framework led to the articulation of several empirical laws within the social sciences and linguistics, the most renowned of which is known as Zipf's Law (Brown, 2005).

Zipf identified a power-law relationship between word frequency and word rank, observing that the most frequent word in a corpus typically occurs approximately twice as often as the second most frequent word, three times as often as the third, and so forth. This empirical pattern, known as the Zipfian distribution, characterizes the frequency $f(r)$ of an item's occurrence in a finite corpus as inversely proportional to its rank r, as given in Eq. (1) (Būdienė and Gruodis, 2012). It actually states that, for a corpus of sufficient size, the frequency of use f of any word decreases with its rank r in the text in an approximately hyperbolic way, i.e. $f \propto 1/r$, where "$\propto$" denotes proportionality. Specifically, r stands for the rank or the frequency-related position of the word in the frequency list, where the most frequent word is assigned $r = 1$, the second most frequent $r = 1$, and so on (Moreno-Sanchez et al., 2015). A general formulation of the Law is given as:

$$f(r) \propto 1/r^{\alpha} \quad (1)$$

where the exponent α is close to 1.

There have been discussions on additional mathematical formulations that account for different languages and incorporate correlations with additional empirical laws, such as e.g. Heaps' law and Hedal law related to the correlation between the distribution of vocabulary size and the text size and frequency of individual items, respectively (Budiene and Gruodis, 2012; Yu et al., 2016).

In this paper, we will remain focused on the Zipf's Law in terms of the frequency-rank ratio related to the frequency word lists.

1.1. Zipf's Law and Word Frequency

In designing language courses, the language instructors pay attention to the words to include and focus on. This is important because various words of a language occur with different frequencies. Typically, the ten most frequent words in a language cover from 25-35% of texts, the first 100 words of English cover around 50%, and the most frequent 1000 words are expected to cover between 70% and 90% of a text (Nation, 2016). This would, of course, depend on the type of text, i.e. the professional domain it serves.

Another key aspect of Zipf's Law is the steep decline in word frequency across a ranked frequency list. By the midpoint of such a list, many words typically occur only once within the corpus. This characteristic underscores the importance of frequency-based word lists in the development of effective language curricula -particularly in the context of technical or domain-specific vocabulary, which serves as a primary predictor of reading comprehension in English for Specific Purposes (ESP) instruction.

While the above generalities of the Zipf's Law has been extensively examined in relation to General English (GE) corpora and other languages – including constructed languages such as Esperanto (Manaris et al., 2006) – its application to technical texts remains comparatively underexplored. Adopting established methodologies for the development of English for Specific Purposes (ESP) word lists, this study seeks to investigate the relevance and applicability of Zipf's Law within the domain of Maritime English. Given that Marine Engineering English has been identified as one of the most lexically demanding technical registers (Hsu, 2014; Đurović et al., 2021), we aimed to analyze patterns of word frequency versus their ranks within a comprehensive corpus of marine engineering instruction books and technical manuals. This objective gives rise to the following research question: *How applicable is the Zipf's Law to a technical text?*

2. METHODOLOGY AND CORPUS

This phase of the research draws upon the findings and results of a previously published Word List of Marine Engineering Instruction Books and Manuals (Đurović, 2021), specifically utilizing the technical corpus of marine engineering instructional texts that was compiled for that study.

The corpus constitutes a comprehensive compilation of marine engineering instructional texts and technical manuals encompassing a wide range of onboard propulsion and auxiliary machinery (Đurović, 2021; Đurović et al., 2021). Following the systematic removal of metadata, typographical errors, and conversion artifacts, the finalized representative corpus comprises approximately 1,769,821 running words. For the sake of convenience, this corpus will hereinafter be referred to as the Marine Engineering Technical Corpus (METC).

In accordance with established methodologies employed in the compilation of numerous lexical databases, we applied a similarly rigorous approach

in the preparation of the corpus. Various stages of the lexical analysis were conducted using software tools developed by linguist Laurence Anthony (2013). Specifically, the AntFileConverter was utilized for the initial transformation of source texts into the requisite plain text format (.txt), followed by the use of AntConc for subsequent lexical and frequency analyses.

3. MARINE ENGINEERING CORPORA AND ZIPF'S LAW

In simplified terms, Zipf's Law posits that the product of a word's rank and its corresponding frequency remains approximately constant; however, in empirical applications, this relationship is rarely observed with absolute precision (Nation, 2016). An idealized representation of Zipf's Law applied to a corpus of ten thousand words is presented in Table 1. This table has been augmented to include their respective coverage of the corpus (expressed as a percentage) to facilitate subsequent comparative analyses.

Seeking to examine the applicability of the Zipf's Law on our highly technical corpus, we tested it through the AntConc software tool (2013; updated versions, 2025) offering the possibility of creating word lists, i.e. their respective ranking in accordance with their frequency in the text. In Figure 1 we are presenting the first 10 most frequent words from the corpus of marine engineering instruction books and manuals, ranked by the AntConc software:

Rank	Freq	Word
1	144202	the
2	43205	of
3	40288	and
4	32771	to
5	26098	in
6	25101	is
7	21247	a
8	19676	for
9	19084	be
10	13977	oil

Figure 1. The most frequent words of METC ranked by AntConc

In Table 2, we give a more detailed overview of the results, with reference to Table 1.

Table 1. The ideal patterning of Zipf's law in a GE text (Nation, 2016)

Rank	Word	Frequency (per 10,000)	%	Rank × frequency
1	the	700	7	700
2	and	350	3.5	700
3	a	233	2.33	699
4	of	175	1.75	700
5	to	140	1.4	700
6	s	117	1.17	702
7	as	100	1	700
700	abide	1	0.01	700
701	absence	1	0.01	701
702	abundant	1	0.01	702
703	accent	1	0.01	703

Table 2. The patterning of Zipf's law in METC

Rank	Word	Frequency (per 10,000)	%	Rank × frequency
1	the	144.202	8.10	144.202
2	of	43.205	2.42	86.410
3	and	40.288	2.26	120.864
4	to	32.771	1.84	131.084
5	in	26.098	1.46	130.490
6	is	25.101	1.41	150.606
7	a	21.247	1.19	148.729

Here we presented only the 7 highest-ranking words, noting that, out of 26,794 word types in our technical corpus, those ranked from 17,303 and on appear only once.

While it is important to acknowledge certain limitations at this stage, it is evident that the product of rank and frequency within our Marine Engineering Technical Corpus (METC) exhibits variability, despite the clearly observable rapid decline in frequency as rank increases. Among the limitations, the specialized technical nature of the texts posed challenges during the scanning and conversion processes, resulting in the removal of a subset of words prior to analysis. Moreover, the corpus-cleaning process is inherently imperfect and non-definitive, such that the final corpus may not fully or precisely represent the complete vocabulary composition and distribution of the source texts.

In light of the aforementioned limitations, we opted for a corpus comprising marine engineering texts that are comparatively less technical and of smaller size, thereby aligning more closely with the characteristics of the corpus employed by Nation (2016). To this end, we constructed a new corpus within the domain of marine engineering that features a more narrative style of discourse. It was derived from *Introduction to Marine Engineering* by David Albert Taylor (2003), totaling 92,808 running words.

After the conversion of our new and smaller corpus of marine engineering, again, using the Anthony's software tools for lexical analyses, we obtained the following results (Figure 2).

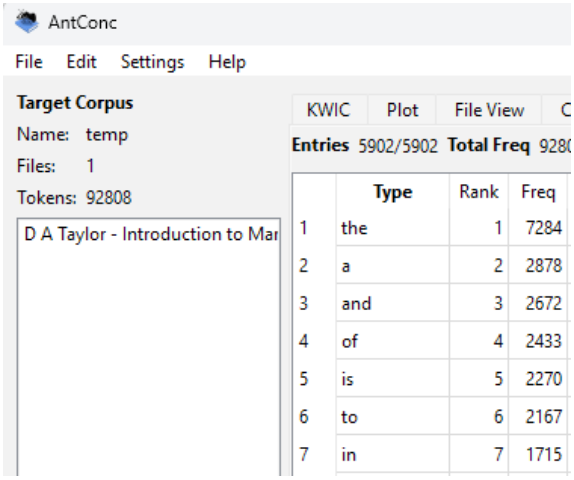

Figure 2. The AntConc results for textbook marine engineering corpus (TMEC)

The results, again, can be presented as those in Table 3.

The comparative findings in the three analyzed corpora are given in Table 4.

As can be seen from Table 4, although vastly different in sizes, the three compared corpora actually show similarities in the text coverage by the most frequent English words. For example, *the* ranges from 7% to 8.10%, *and* from 2.26% to 3.5% etc. Also, both techni-

Table 3. The patterning of Zipf's law in a textbook marine engineering corpus

Rank	Word	Frequency (per 10,000)	%	Rank × frequency
1	the	7,284	7.85	7,284
2	a	2,878	3.10	5,756
3	and	2,672	2.87	8,016
4	of	2,433	2.62	9,732
5	is	2,270	2.45	11,350
6	to	2,167	2.33	13,002
7	in	1,715	1.85	12,005

Table 4. The comparative results for the most frequent GE words coverage

Rank	10,000 GE (Nation)	%	METC	%	Textbook MEC	%
1	the	7	the	8.10	the	7.85
2	and	3.5	of	2.42	a	3.10
3	a	2.33	and	2.26	and	2.87
4	of	1.75	to	1.84	of	2.62
5	to	1.4	in	1.46	is	2.45
6	is	1.17	is	1.41	to	2.33
7	as	1	a	1.19	in	1.85

cal corpora have *in* among the first 7 most frequent words, while in GE texts it is *as*. The indefinite article also shows somewhat reduced occurrence in the most technical text of marine engineering instruction books, which can speak in favor of the technical lexicon, but also syntax of predominantly technical texts.

It is also observable that word frequency tends to decline more gradually in technical corpora compared to general English (GE) corpora. This phenomenon is consistent with the nature of technical texts, which typically concentrate on a relatively narrow professional domain and consequently encompass a more limited range of lexical items. In such contexts, linguistic richness and vocabulary diversity are often deprioritized, if not constrained. It is important to note that the comparison is made against an idealized pattern, whereas empirical linguistic data rarely exhibit the same degree of conformity to Zipf's Law.

It is important to emphasize the distinction between different types of lexical units when compiling and analyzing word lists. Typically, word lists are presented as headwords, whereas lexical analyses often expand these to include all members of the corresponding word families. In contrast, the AntConc software generates results based on word types. In the context of our most frequent English word list, this distinction has minimal impact, as the list is predominantly composed of function words (e.g., *the*, *a*, *and*, *of*, *is*).

Additionally, this study focused primarily on the most frequent English words across various text types; however, future research could extend the analysis to examine the frequency and ranking of mid-frequency/technical vocabulary.

4. CONCLUSION

Motivated by the importance of technical vocabulary frequency in English for Specific Purposes (ESP) course design, as well as by general linguistic regularities and patterns in language use, this study sought to evaluate the applicability of Zipf's Law within marine engineering technical texts. Drawing upon the simplified formulation and findings of Paul Nation concerning the most frequent general English (GE) words and texts, we conducted a comparative analysis of two categories of marine engineering texts, broadly classified as more and less technical in nature. The results, however, revealed no significant differences between these categories, while both exhibited a marked deviation from the constant values predicted by Zipf's Law in natural or GE texts. Nonetheless, certain features of the law – such as the steep decline in frequency and the coverage of the most frequent GE words – appear to persist across diverse English text types, including highly technical corpora.

The above brings us back to the initial definition of the Law as e one representing a universal distribution for natural language (not depending on chosen language, word quantity and specific sphere of language use (Budiene and Gruodis, 2012). This, again, points to the difference between natural and technical language, although the borders can be rather vague. In addition, abundant GE corpora (e.g. BNC, COCA, Brown, FLOB etc.) include also texts on specific areas of human activities and interests.

The foregoing discussion returns us to the original conceptualization of Zipf's Law as a universal distribution governing natural language, irrespective of the specific language or corpus size (Budiene and Gruodis, 2012). This universality highlights the distinction between natural and technical language varieties, although the boundaries between these categories may often be fluid and ambiguous. Furthermore, extensive general English corpora, such as the British National Corpus (BNC), the Corpus of Contemporary American English (COCA), the Brown Corpus, and the Freiburg-LOB Corpus (FLOB), encompass texts from a broad spectrum of specialized domains and areas of human activity.

In general, certain core characteristics, such as the respective coverage provided by the most frequent English words and the rapid decline in word frequency with the rise in ran, appear to be consistent across all text types. However, given the documented linguistic particularities of highly technical and professional discourse, English for Specific Purposes (ESP) retains its inherently specialized and profession-oriented nature. As such, each ESP domain should be approached and analyzed in accordance with its distinct lexical and functional demands.

Conflict of interest: The author declares no conflict of interest related to the content of this paper.

REFERENCES

Brown, K., 2005. Encyclopedia of Language and Linguistics, 2nd edition, Elsevier Ltd. Available at: https://www.sciencedaily.com/releases/2017/08/170810082147.htm.

Būdienė, G. & Gruodis, A., 2012. Zipf and Related Scaling Laws. 2. Literature Overview of Applications in Linguistics, Innovative Infotechnologies for Science, Business and Education, ISSN 2029-1035 – Vol. 1(12): 17-26. Available at: http://journal.kolegija.lt/iitsbe/2012/Budiene-zipf-2-IITSBE-2012-1(12)-17-26.pdf.

Đurović, Z., 2021. Corpus Linguistics Methods for Building ESP Word Lists, Glossaries and Dictionaries on the Example of Marine Engineering Word List, Lexikos 31. Available at: https://doi.org/10.5788/31-1-1647.

Đurović, Z., Vuković Stamatović, M. & Vukičević, M., 2021. How Much and What Kind of Vocabulary do Marine Engineers Need for Adequate Comprehension of Ship Instruction Books and Manuals?. Circulo de Linguistica Aplicada a la Comunicacion, 88: 123–133. Available at: https://dx.doi.org/10.5209/clac.78300.

Hsu, W., 2014. Measuring the Vocabulary Load of Engineering Textbooks for EFL Undergraduates, English for Specific Purposes 33: 54–65. Available at: https://doi.org/10.1016/j.esp.2013.07.001.

https://www.laurenceanthony.net/, 2012. https://www.laurenceanthony.net/software/antfileconverter/, last accessed on 20 June, 2025.

https://www.laurenceanthony.net/, 2012. https://www.laurenceanthony.net/software/antconc/, last accessed on 20 June, 2025.

Manaris, B. et al., 2006. Investigating Esperanto's Statistical Proportions Relative to other Languages using Neural Networks and Zipf's Law, Proceedings of the 2006 IASTED International Conference on ARTIFICIAL INTELLIGENCE AND APPLICATIONS (AIA 2006), February 13 – 16, 2006, Innsbruck, Austria. Available at: https://bpb-us-w2.wpmucdn.com/blogs.charleston.edu/dist/e/813/files/2020/10/IASTED2006.pdf.

Moreno-Sanches, I., Font-Clos F. & Corral, A., 2015. Large-scale analysis of Zipf's law in English texts, PLOS One, 11(1): e0147073. Available at: https://doi.org/10.1371/journal.pone.0147073.

Nation, I.S.P., 2016. Building and Using Word Lists for Language Learning and Testing, 1st Edition, Amsterdam/Philadelphia: John Benjamins Publishing Company. Available at: https://doi.org/10.1075/z.208.

Taylor, D.A., 2003. Introduction to Marine Engineering, Revised 2nd Edition, Oxford: Elsevier. Available at: https://shop.elsevier.com/books/introduction-to-marine-engineering/taylor/978-0-7506-2530-2.

Yu, S., Xu, C. & Liu, H., 2018. Zipf's Law in 50 Languages: Its structural pattern, linguistic interpretation, and cognitive motivation. Available at: https://doi.org/10.48550/arXiv.1807.01855.