9th International scientific conference
Technics and Informatics in Education – TIE 2022
16-18 September 2022

Session: *ESP and Technics and Informatics: Challenges and Perspectives*
Original scientific paper
DOI: 10.46793/TIE22.379Q

Key ESP Words and Phrases

Zorica Đurović
University of Montenegro, Faculty of Maritime Studies Kotor, Montenegro
zoricag@ucg.ac.me

Abstract: *We frequently mention or use the concept of keywords. However, many are unaware of the possibilities of dealing with them on a different level, which would include accurate statistics and a justified selection of words and phrases (n-grams) that can be considered specific vocabulary and word clusters for a certain type of text. The paper aims to present a possible and reliable method of providing such lexical information for specific technical genres. In our case, it would be a collection of marine engineering technical manuals for tanker ships. The purpose of the methodology presented is to provide a lexical tool that can be applied to any technical genre or more of them and that provides us with useful and concentrated ESP vocabulary material to be used in ESP classes and courses.*

Keywords: *keywords; n-grams; marine engineering; technical manuals.*

1. INTRODUCTION

Technical vocabulary is what the specifics of a Language for Specific Purposes (LSP) mostly pertain to. Therefore, special attention is always paid to the selection and design of vocabulary teaching material. On the assumption that our language learners have mastered the basics of their target non-native language, the main idea is to provide them with "early specialization" in their professional language [1]. New momentum in vocabulary research has been brought by information technology assets enabling practical and fast collection and creation of electronic corpora, along with computer software solutions for lexical analysis of texts. This has provided us with the opportunity to gain more accurate statistical analysis and justification of data used in vocabulary analysis and teaching material. The concept of keywords, for example, has been widely used and applied, as implied by the meanwhile created and parallelly used compound (*keywords*). Here, however, we present the possibility of statistically justified, software-based extraction of key words (or *keywords*) and phrases (word clusters) from a technical genre.

1.1. Target language learners

The syntagm *language learners* here and, as usual, does not necessarily refer to those studying a foreign language, but generally to non-native speakers who need the second language to accommodate the professional discourse community they belong to. One of the most effective examples is the maritime community sharing English as their *lingua franca* around the globe. Maritime English comprises many different registers and communicative purposes. In this paper, we deal with English for Marine Engineering Purposes, which proved to be one of the most demanding ESPs vocabulary-wise [2][3]. Our target language learners are therefore the students of Marine Engineering and active seafarers during their lifelong learning process and particular courses they undergo during their professional careers.

1.2. Corpus

One of the main professional tools of marine engineers, once they sign on vessels, are ship's instruction books and manuals. They are indispensable in familiarizing with the ship's systems and devices, as well as for their regular maintenance, repairs, and overhauls. Having in mind the current and prospective trends in shipping, we opted for technical manuals of tanker ships. Following the expert advice, we provided a comprehensive selection of 61 technical manuals from a modern tanker ship. Due to practical reasons, as well as to avoid the commercialization of the data, we are not presenting the corpus selection in more detail. In general, the Corpus of Tanker Ship Technical Manuals (CTSTM) contains instruction books and manuals for the main engine, generators, lubrication system, separator, economizer, incinerator, sterilizer, valves, steering gear, shafting, condenser, filters, pumps, and other auxiliaries, gears, and systems. In total, the corpus amounts to 1,109,080 running words or tokens, obtained after an attentive "cleaning" and preparation of the corpus for further analysis.

2. METHODOLOGY

The intention of this paper is to present a methodology that can provide us with statistically accurate lexical information on a type of text. It is shown on the example of a markedly technical and actual type of marine engineering publications (CTSTM). It aims to tackle the demand of such a text vocabulary-wise, as well as to provide a recommendation for the extraction of words and phrases (n-grams) that can justifiably be considered key for the particular text or genre.

To investigate the lexical profile and demand of the target Corpus of Tanker Ship Technical Manuals, we used the freeware tool AntWordProfiler, version 2.0.1. [4]. To accommodate the software requirements, the *.pdf* files were converted to the *.txt* format (plain text). The referent General English (GE) word lists used for the process were the Nation's word lists produced from the British National Corpus and Corpus of Contemporary American English (BNC/COCA). These 25 lists contain about 1,000 word families[1] each, and, for this kind of research, they are usually accompanied by additional lists of the most frequent proper names, abbreviations, transparent compounds, and marginal words [5][6][7].

For keywords specifically, we used AntConc, version 4.1.0. by the same developer [8]. This software provides us with the opportunity to obtain the list of corpus keywords, comprised of the words unusually frequent as compared to a referent corpus of General English (GE). As such, these words are considered to reflect the nature of a text or genre and enable its better and proper comprehension [9]. As for the referent GE corpus, we used the Freiburg-Lancaster-Oslo/Bergen Corpus (FLOB). This GE corpus was developed aiming to produce a contemporary British English corpus serving as a counterpart of the Brown University Standard Corpus of Present-Day American English [10].

In addition, we used the same software to examine then-grams or multi-word units most frequently occurring in this specific professional genre and therefore worthwhile pursuing.

3. LEXICAL PROFILE OF CTSTM

Firstly, we wanted to examine the lexical profile and demand of our target corpus, thus we tested it against the GE word lists (BNC/COCA) as per the methodology given above.

Having in mind the findings and agreement of relevant authors of the area that adequate reading comprehension is expected at the level of 95% of known vocabulary [11], we can see that in our target corpus it is not reached even with all the available 25,000 General English words2, not to mention the ideal threshold of 98% [12].

Table 1. *Coverage of GE word lists in TSTM*

BNC/COCA Word Lists	Coverage %
2,000 + proper names, abbreviations, compounds and marginal words	61.54
3,000 + proper names, abbreviations, compounds and marginal words	85.32
4,000 + proper names, abbreviations, compounds and marginal words	88.25
5,000 + proper names, abbreviations, compounds and marginal words	90.39
6,000 + proper names, abbreviations, compounds and marginal words	91.17
7,000 + proper names, abbreviations, compounds and marginal words	91.9
8,000 + proper names, abbreviations, compounds and marginal words	92.46
25,000 + proper names, abbreviations, compounds and marginal words	94.25

If we take into consideration that about 4,000 GE words are considered sufficient for adequate reading and understanding of, for example, newspapers [13] or for successful listening and understanding of academic lectures and TED talks related to physics [14], or that as many as 12,000 are needed for some highly professional genres [3], the results point to the challenges imposed by the technical nature of our target corpus of tanker ship technical manuals. Taking into account the recommendations for early language specialization when it comes to ESP [1], our aim here is to explore the most frequent keywords and phrases found in technical manuals meant for marine engineers on tanker ships.

4. KEY WORDS IN CTSTM

Unlike the frequency counts, the keyness of a word does not necessarily anticipate a high but rather unusual frequency of that word as compared to its use in the general language, in our case – General English. Keywords are consequently those with a "special status" [15] in a genre, reflecting its specificity when compared to other types of texts. The tools enabling us to relatively easily extract keywords from a text or corpus especially come in handy, providing us with meticulously organized lexical and syntactical material [7].

The initial and total keyword list counted 92 lemmas. The keyness in our approach, however, does not refer to an individual lemma, as presented by the software. Lead by the principle of learning burden or effort put in mastering a word [16], we put and counted together word family members, adding the members to the one with the highest frequency. That way we added, e.g., *setting* to the *set* "family", *cleaning* to *clean*, *operating* to *operation,* and similar (Table 2), adding their

[1] A word family includes the head or base word with all its inflected and derived forms.

[2] A word here denotes a word family.

keyness and frequency values, as well. We also excluded the most frequent English words from the list such as: *if*, *be*, *is*, which mostly belong to the 10 most frequent words of the English Language [17][18]. Also, regardless of our best efforts to remove proper names, single letters, symbols, and abbreviations from the initial corpus, some still occurred in the list, so we removed those as well. Finally, we are presenting the list of 78keywords in CTSTM, arranged by their cumulative frequency ranging from +2 to +8,906, as per the previously explained process (Table 2).

Table 2. *Key words in CTSTM*

No.	Word	No.	Word
1	oil	40	position
2	valve, valves	41	must
3	pressure	42	fig
4	pump	43	load
5	operate	44	ring
6	control	45	language
7	water	46	output
8	step	47	screw
9	check	48	actuator
10	air	49	compressor
11	separator	50	signal
12	manual	51	maintenance
13	boiler	52	bar
14	speed	53	level
15	burner	54	replace
16	fuel	55	sensor
17	system	56	instructions
18	start	57	note
19	set, setting	58	remove
20	engine	59	hydraulic
21	unit	60	supply
22	motor	61	piston
23	temperature	62	shaft
24	installation	63	bearing
25	type	64	cable
26	stop	65	page
27	safety	66	filter
28	alarm	67	cylinder
29	flow	68	turbocharger
30	mode	69	feed
31	parts	70	clean, cleaning
32	menu	71	terminal
33	bowl	72	data
34	input	73	shut
35	governor	74	service
36	switch	75	gasket
37	figure	76	inlet
38	panel	77	value
39	steam	78	spindle

For practical reasons, we are not presenting additional data such as respective keyness and frequency counts. Nevertheless, for illustrative purposes, we are giving a shortened overview of data obtained through the software in Table 3. The example covers the highest-ranked keywords in the corpus:

Table 3. *The five highest-ranked keywords in CTSTM*

Rank	Word	Keyness	Frequency
1	oil	+8,906	8984
2	valve	+7,113	7,120
3	operation	+6,610	6,766
3	pressure	+5,164	5,796
4	pump	+ 4,907	4,921
5	operation	+4,378	4,492

As we have a closer look at the composition of the keyword list (Table 2), we can see that most of the words reflect the specificity of the marine engineering lexicon, especially that of a tanker ship, such as *oil*, *valve*, *operation*, *pressure*, *pump*, *maintenance* and similar. However, we also come across a few notions that belong to other or general registers, such as e.g. *language*. Being curious about the unusual frequency of the word language in a highly technical genre, we explored another software advantage referring to collocations. We found out that the word *language* here frequently collocates with *selection*, *English*, *menu*, *table*, *on,* etc. Its frequency is therefore explained by instructions on settings the corpus is abundant with. A similar examination can be done for any of the words, seeking their collocations or word clusters, which can be of additional use to material and course designers, as well as for the language learners themselves.

5. THE MOST FREQUENT N-GRAMS IN CTSTM

Bearing in mind that word semantics is context-dependent, our further interest would be driven towards the most common combinations of words we can come across in this specific type of manual. For this purpose, we sought to detect the most frequent n-grams consisting of 2–5 members (words). The examples presented in Table 4 are the most frequent ones with each cluster including either (at least) two nouns, an adjective and a noun or a verb and a noun, in order to avoid the most frequent n-grams in general language, such as *of the*, *to the*, etc. and also to pursue the examples of the most frequent collocations in the corpus. Again, since the software provides lemmatized results, we put together similar expressions, including those with additional prepositions and/or articles (e.g. *(if) this is not the case*). With additional content words, we retained a separate count (e.g. *direction of rotation* and *check the direction of rotation*).

Interestingly, there were no distinctive 3-grams in the final list (Table 4).

Table 4. *Most frequent n-grams in CTSTM*

No.	N-grams	Frequency
	5-grams	
	in such a way that	62
	(if) this is not the case	50
	it is not possible to	40
	it is recommended that the	24
	the serial number of the	18
	attention must be paid to	16
	work must be carried out	12
	4-grams	
	as described/shown in chapter/section/figure	144
	(check) the direction of rotation	94
	failure to comply with	34
	the first start up	34
	check the oil level	32
	from time to time	32
	state of the art	26
	2-grams	
	fuel oil	1,366
	control system	1,204
	oil pump	778
	spare parts	691
	operation manual	600
	oil flow	576
	solenoid valve	550
	set point	504
	safety valve	486
	data sheet	480
	stop valve	460
	control unit	450
	compressed air	438
	control valve	432
	oil pressure	430
	technical data	422

6. GE PHRASES IN CTSTM

However professionally and technically oriented, English courses, naturally, cannot be strictly focused on the technical vocabulary, but must also be accompanied by General English skills, adapted to the practical needs of our language learners. As we could see in Sections 4 and 5, we were seeking primarily technical collocations and word clusters typical (or key) for our target professional corpus. However, the above-described and applied methodology can greatly assist language teachers in extracting the most frequent GE phrases worth focusing on (Table 5). Additional exercises can then be developed to help language learners master them in terms of productive language skills.

Table 5. *The most frequent GE phrases in CTSTM*

No.	Phrase	Frequency
1	by means of	1072
2	in (the/this) case (of)	588
3	(should) be carried out	512
4	in order to	416
5	in accordance with	316
6	as well as	274
7	make sure that	138
8	it is recommended	114
9	care should be taken	106
10	in such a way	86

7. CONCLUSION

When selecting and organizing vocabulary teaching material for language learners, different approaches, more or less deliberate, are applied. It is of particular importance and challenge when it comes to ESP, such as, in our case, a very specific English for Marine Engineering Purposes. We, therefore, presented a software-based corpus linguistic method for the extraction of target or key technical vocabulary, as well as the most frequent word clusters. From the technical corpus of 1,109,080 tokens, we obtained the keyword list of 78 words (word families) with an additional list of 2—5-grams. In addition, we used the same methodology to elicit the frequency list of GE phrases most frequently found in our target corpus of tanker ship technical manuals. The methodology presented is replicable in the case of any ESP and can be of assistance to both language teachers and learners. Special attention, however, should be paid to each step of the process and its justification, from the proper selection and preparation of representative corpus, through the software settings and operation, to the organization of the final results, their adaptation, and proper use. Above all, the practical and professional needs of our language learners should be born in mind throughout the process.

REFERENCES

[1] Coxhead, A. & Hirsch, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliqueé, 12(2),* 65–78.

[2] Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes, 33*, 54-65.

[3] 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. https://dx.doi.org/10.5209/clac.78300

[4] Anthony, L. (2022). AntWordProfiler (Version 2.0.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

[5] Nation, I. S. P. (2004). A Study of the most frequent word families in the British National Corpus. In P. Bogaards and B. Laufer (Eds.) *Vocabulary in a second language, Selection, acquisition and testing*. Amsterdam: John Benjamins, 3–13.

[6] Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63(1),* 59–82.

[7] Đurović, Z., Nicolas, C. and Jurkovič, V. (2022). Keywords and phrases in technical manuals on oil spills. 20th International Conference on Transport Science ICTS 2022, Portorož, Slovenia, 23-24 May, *Conference Proceedings,*107-113. https://icts.sdzp.org/wp/wp-content/uploads/2022/06/ICTS-2022-Proceedings-CIP.pdf

[8] Anthony, L. (2022). AntConc (Version 4.1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

[9] Culpeper, J. & Demmen, J. (2015). Keywords. In Biber, D. & Reppen, R. (eds.) *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press. DOI: 10.1007/9781139764377.006

[10]Kucera, H. & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: R. I. Brown University Press.

[11]Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In *Special language, from humans thinking to thinking machines*, Ed. Lauren, C. and Nordman. M. Multilingual Matters, 316–323.

[12]Hu, M. & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a foreign language, 23*, 403–430

[13]Nation, I.S.P. (2006). "How large a vocabulary is needed for reading and listening?" *Canadian Modern Language Review, 63(1),* 59-82.

[14]Vuković Stamatović, M. (2019). Vocabulary complexity and reading and listening comprehension of various physics genres, *Corpus Linguistics and Linguistic Theory*. DOI: https://doi.org/10.1515/cllt-2019-0022

[15]Stubbs, M. (2010). Three concepts of keyness. In M. Bondi, and M. Scott, (Eds.) *Keyness in texts*. Amsterdam: John Benjamins.

[16]Nation, I. S. P. (2000). Review of what's in a word? Vocabulary development in multilingual classrooms by N. McWilliam, *Studies in Second Language Acquisition, 22(1),* 126–127.

[17]Nation, I. S. P. (2013). *Learning vocabulary in another language* (second edition). Cambridge: Cambridge University Press.

[18]Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins.