8th International Scientific Conference
Technics and Informatics in Education
Faculty of Technical Sciences, Čačak, Serbia, 18-20th September 2020

Session *Teacher Professional Development and General Education Topics*

Professional paper
UDC: 811.111:004.738.5

Some Corpus Linguistics Tools for Adequate Reading Comprehension of Instruction Books for Electro Technical Officers

Zorica Đurović [1*] and Nikola Marvučić [1]
[1] University of Montenegro/Faculty of Maritime Studies Kotor, Kotor, Montenegro
[*] zoricag@ucg.ac.me

Abstract: *The development and arising of new Englishes for specific purposes imposes great challenges to language teachers and course designers. Nevertheless, the ever boosting information technologies bring new possibilities in terms of abundant corpora collection and analysis and production of relevant and profession-oriented course materials. Considering this specific research, contemporary corpus linguistics tools enabled us to obtain precise results on the vocabulary load and complexity of instruction books for Electro Technical Officers, aiming to reach the adequate or ideal threshold of reading comprehension. The paper results provide comparative analyses, leading to a clear call for a word list creation to suit the vocabulary needs of Electro Technical Officers in the most efficient way.*

Keywords: *reading comprehension, corpus, word list, instruction books*

1. INTRODUCTION

The technological development after the World War II influenced all areas of human endeavors. The new era of informatics brought new possibilities to the regular life of people, but especially to all fields of science. Following the accelerated development of cross-border scientific and business activities, the need for a common language and the economic and military superiority of the USA have imposed English as the *lingua franca* of the modern world. This especially became the case in maritime affairs, formally effective as of the establishment of International Maritime Organization in London in 1948.

The ever growing and multiplying scientific and business activities have been generating a multitude of branches and sub-branches of English for Specific Purposes (ESP). This has imposed new challenges and requirements on English language teachers in their effort to develop course materials and methodologies according to the needs of the (English) language learners.

Fortunately, the development of information technology has also provided for new tools and methodologies, such as corpus linguistics methods and software, enabling easier access to the abundance of real-life material and its multiple analysis.

2. THEORETICAL BACKGROUND

Although corpus linguistics methods had been known and applied earlier, no sooner than in the 60s of the previous century, the new array of linguistic research was opened, owing to new possibilities brought by information technologies. This has brought many benefits to language teachers and course designers, and consequently the language learners as well.

Bearing in mind the fact that specific, technical vocabulary is considered the main meaning predictor in reading comprehension, the renewed interest in vocabulary has led to the development of modern computer tools for lexical analysis of written or spoken texts. The programs and software such as AntWordProfiler, AntConc, Wordsmith, Sketch Engine, LancsBox and many others, have been used to provide data on the quantity and type of vocabulary and its various elements and patterns in a certain text type (genre). In addition, some of them can be used for extracting key words and word lists according to word frequency values. The idea behind this kind of analyses and produced word lists is to achieve the adequate reading comprehension with less learning effort, which would be especially useful with our target learners.

2.1. Reading Comprehension

In trying to answer the question on the quantity of needed vocabulary, two thresholds have been mostly used in the lexical profiling of texts. The level of 95% of known words in a text was set as desirable by Laufer [1], while Nation [2] advocates for 98% of known vocabulary for achieving appropriate reading comprehension. Thus, the threshold of 95% has been used as the adequate,

and 98% as the ideal coverage of a text, whilst the remaining 2–5% are expected to be understood from the very context.

3. METHODOLOGY AND CORPUS

The method we used for our corpus/vocabulary analysis and research is called Lexical Frequency Profiling [3]. It will enable us to answer the question on the quantity of certain types of vocabulary within our referent corpus. The software that we have found most useful and convenient for this purpose is AntWordProfiler 1.4.0w developed by Laurence Anthony [4], as an upgraded version of the previously used RANGE program [5].

In our analysis, we used the relevant and available word lists of general English (GE), academic English and ESP. An additional program – Familizer + Lemmatizer [6] was used to expand the available word lists into all-family-members form, so that they can be used by the software for lexical profiling of the text. Word family is the most commonly used unit of measurement in this type of research, meaning that a "word" would anticipate the headword with all its derivatives and inflected forms.

For the preparation and conversion of the corpus into the "plain text" format, we used AntFileConverter [7], after which we additionally "cleaned" the text from tables, lists, references, typos and graphical errors, in order to make the analysis as precise as possible. These efforts can often be very demanding and time-consuming, but are still very important for the relevance of the results obtained [8].

3.1 Word lists

Following the previous research of the area, we tested our corpus against the most used general English word lists, the academic word list and available and relevant ESP word lists.

General Service List (GSL) was produced by Michael West in 1953 [9]. It comprises 2,000 most frequent words from a 5-million-word corpus of the English Language. It is far from a contemporary one, but it is still commonly used for general English coverage testing, especially with Academic Word List [10] which was built upon it. They have been used together in this kind of analysis, so, we are going to use them in the same way for comparison reasons.

Concerning more contemporary general English, Paul Nation extracted 25 general English word list (each of 1,000 word families), with additional four covering some most frequent proper names, abbreviations, transparent compounds and marginal words [11]. This set of GE words (known as BNC/COCA word lists) was derived from the British National Corpus (cc.100 million words) and Corpus of Contemporary American English (cc. 450 million words).

After mastering the first 2,000–3,000 general (English) words, it is recommended that language teaching/learning should be more field oriented, the first step towards specialized word lists would be an academic word list of vocabulary common to various scientific and academic fields. The most used one is the Academic Word List by Averil Coxhead [10], comprised of 570 word families and expected to cover about 10 % of academic texts.

More and more specialized word lists have been produced in order to achieve significant coverage in relevant texts and adequate reading comprehension in the most efficient and least time-consuming way. To the best of our knowledge, only two available word lists of those [12] [13] developed to this date are applicable to our research, considering the scientific field, student level and software requirements. Both Ward's and Hsu's lists belong to more general engineering English and have been extracted from undergraduate textbooks (25 and 100 of them, respectively) for engineering students of various engineering fields.

3.2 Corpus

Our target (English) language learners are the students of marine electrical engineering, but also seafarers undergoing trainings for Electro Technical Officer (ETO). Considering their most practical language needs in terms of their onboard duties and operations, we came to the conclusion that adequate reading comprehension of ship's instruction books are of utmost importance to their daily duties. In particular, from the day of signing on, all marine engineers need these "tools" to familiarize themselves with the ship systems and devices, and their daily duties anticipate regular maintenance and monitoring, but also repairs and overhauling as required. This is especially relevant for cruise ships where Electro Technical Officers are more numerous than on other types of vessels, having a hierarchy of their own. Following this fact and some expert advice, we collected instruction books for a very common and popular Voyager class of cruise ships, selecting their contents with reference to electrical installations, maintenance and repairs. This class of vessels has the carrying capacity of 3,800–4,000 persons and the gross tonnage of about 138,000. Bearing in mind the systems and devices maintained by ETOs, the corpus comprises instruction books for diesel-electric propulsion, power generation and distribution systems, various automation systems, switchboards, dynamic positioning systems, engine room machinery (electrical part), as well as those related to cooling and ventilation systems, steering equipment, signaling, radio and safety equipment, lighting and other numerous ancillary services. The

final corpus amounts to 176,491 running words (tokens) extracted from 44 electronic (scanned) files of varying volume.

4. RESULTS

In order to test the vocabulary types in our referent corpus, we first analyzed it against the General Service List [9] and Academic Word List [10], 2000), as these are usually used together in this kind of research. The obtained results (Table 1.) show a significantly lower GSL coverage (66.95%) compared to the usual coverage of 78 – 98% reported for various types of written texts [14], and somewhat closer coverage to that in various academic texts of up to 76% [10], which points to the extremely technical nature of this professional type of texts.

Table 1. *The coverage of GSL and AWL in the corpus of instruction books for ETOs*

Word lists	Tokens	Coverage (%)
GSL	118,165	66.95
AWL	12,898	7.31
-	45,428	25.74
Total	176,491	100

The coverage of AWL is also lower than the average of 10% in various academic texts [10]. In other words, knowing the first 2,000 general English words and the most frequent academic words would leave the reader with about a quarter of the unknown vocabulary, or every fifth word unknown, which would make the reading and understanding of this type of text extremely difficult.

For the purpose of illustration, we can present another possibility of the AntWordProfiler software. In the excerpt below (Figure 1.), vocabulary types are given in different colors, i.e. the words marked red belong to the first 1,000 GSL words, the second 1,000 are marked green, whilst the words colored in red belong to AWL. More colors are available for additional lists, whilst the words remaining outside the referent lists remain black.

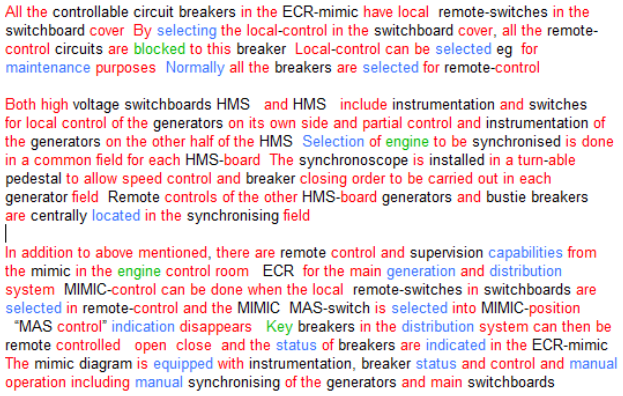

Figure 1. *A corpus excerpt with GSL and AWL coverage given in colors*

Aiming to determine the specific amount of general English vocabulary needed to reach the adequate reading comprehension (95 – 98%) we tested our corpus against the BNC/COCA word lists [11], including the additional lists of the most frequent proper nouns, abbreviations, marginal words and transparent compounds.

Table 2. *The coverage of the BNC/COCA lists in the corpus of instruction books for ETOs*

BNC/COCA Word lists	Tokens
2,000 + proper n., *abbrev. and marginal words*	77.3
3,000 + proper n., *abbrev. and marginal words*	85.48
4,000 + proper n., *abbrev. and marginal words*	88.5
5,000 + proper n., *abbrev. and marginal words*	90.71
6,000 + proper n., *abbrev. and marginal words*	91.15
7,000 + proper n., *abbrev. and marginal words*	91.56
8,000 + proper n., *abbrev. and marginal words*	91.97
25,000 + proper n., *abbrev. and marginal words*	93.58

The results (Table 2.) show us that the first 2,000 general English words (BNC/COCA) cover 77.3% of our corpus, which is significantly higher than GSL coverage (66.95%). This is understandable, considering that the BNC/COCA word lists are much more up-to-date in terms of their age and selection of texts, as well as the contemporary nature of our referent corpus. Another advantage of the letter type of general English word lists is that they can show us higher levels of GE coverage. In our case, we can see that not even the adequate threshold of 95% of needed vocabulary is reached with general English words, not to mention the ideal coverage of 98%. This would mean that this kind of texts would be a hardly attainable task even to a native speaker not involved in (marine) electrical engineering.

As illustrated before, the classification of the vocabulary by referent lists can be presented with the aid of different colors. On the example of the same short excerpt, the first 1,000 is given in red, the second 1,000 in green, the third 1,000 in blue, the fourth 1,000 in pink, the fifth 1,000 in violet, the sixth 1,000 in orange, the seventh 1,000 in brown, the eight 1,000 in dark blue, etc., and the remaining vocabulary is given in black.

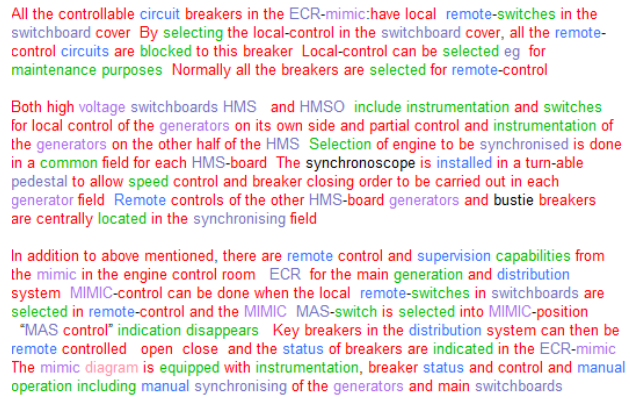

Figure 2. *A corpus excerpt with BNC/COCA coverage given in colors*

Not being able to reach the adequate reading comprehension with general English words only, we tested the coverage of available and applicable engineering word lists.

The coverage of the Ward's basic engineering English word list (BEEWL) is somewhat lower than in his original corpus of engineering textbooks (16.4%), but is still substantial and could be of good use to marine (electrical) engineers, as well (Table 3.). In our analysis we used it without the first 2,000 most frequent GE words, since Ward did it the same way due to the poor vocabulary skills of his students.

Table 3. The *coverage of BEEWL in the corpus of instruction books for ETOs*

Word lists	Tokens	Coverage (%)
BEEWL (Ward)	25,550	14.48

Hsu, however, first excluded the most frequent English words in her analysis, so we followed the same methodology (Table 4.). The coverage of her engineering English word list (EEWL) is significantly below the coverage in her original corpus of textbooks from 20 various engineering areas (14.3%).

Table 4. *The coverage of EEWL in the corpus of instruction books for ETOs*

Word lists	Tokens	Coverage (%)
BNC/COCA 2,000	120,685	68.38
EEWL (Hsu)	16,677	9.45
Total	137,362	77.83

5. CONCLUSION

Using some of the most contemporary corpus linguistics software and programs, we were able to provide data on vocabulary type and load of our referent corpus. Not surprisingly, the results point to the extremely technical and demanding nature of the texts vocabulary-wise, the adequate reading comprehension of which cannot be reached with knowing general English vocabulary only. This justifies the general recommendation of upgrading the first several thousands of general English words with more specific and profession-oriented vocabulary, which is a general tendency in English for specific purposes.

The fact that the adequate threshold has not been reached even with the existing word lists of adjacent engineering fields, clearly calls for the creation of a specific ETO word list. Fortunately, the software used here provides us with the very possibility, thus the authors hope to present the target learners, as well as other course designers and authors, with this kind of ESP aid in the near future.

REFERENCES

[1] Laufer, B. (1992). What percentage of text-lexis is essential for comprehension? In *Special Language: From Humans Thinking to Thinking Machines*, Ed. Lauren. Ch. And M. Nordman, Multilingual Matters, 316–323.

[2] *Nation,* I.S.P. (2006). How Large a Vocabulary Is Needed for Reading and Listening? *Canadian Modern Language Review*, 63 (1), 59–82.

[3] Laufer B. & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, 16(3), 307–322.

[4] Anthony, L. (2014). AntWordProfiler, Build 1.4.1.0., Center for English Language Education in Science and Engineering, School of Engineering, Waseda University, Tokyo, Japan

[5] Nation, I. S. P. and Heatley, A. (1994). Range: A program for the analysis of vocabulary in texts (software), retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx.

[6] Cobb, T. (2018). From corpus to CALL: The use of technology in teaching and learning formulaic language, in A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.) *Understanding Formulaic Language: A Second Language Acquisition Perspective* (pp. 192–211).

[7] Anthony, L. (2017), AntFileConverter (Version 1.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

[8] West, M. (1953). A General Service List of English Words. London: Longman, Green and Co.

[9] Coxhead, A. (2000). A New Academic Word List. *TESOL* 34(2), 213–238

[10] Nation, I. S. P. (2012). The BNC/COCA word family lists. http://www.victoria.ac.nz/lals/about/staff/paul-nation (accessed January 2019).

[11]Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, 28(3), 170-182.

[12]Hsu, W. (2014). Measuring the Vocabulary Load of Engineering Textbooks for EFL Undergraduates, *English for Specific Purposes* 33, 54–65

[13]Waring, R. & Nation, I. S. P. (1997). Vocabulary size, text coverage, and word lists. In *Vocabulary: description, acquisition and pedagogy*. N. Schmitt and M. McCarthy (eds.), Cambridge: Cambridge University Press.