378.147::811.111
811.111’276.6:001
https://doi.org/10.18485/bells90.2020.1.ch26

Milica Vuković Stamatović[*]
Faculty of Philology
University of Montenegro
Montenegro

VOCABULARY SUITABILITY OF POPULAR SCIENCE BOOKS FOR ENGLISH FOR SCIENCE CLASSES: A CORPUS-BASED STUDY

Abstract

English for Science teaching materials and resources are much scarcer than those used for the general English classes. In the light of this, this study aims to explore how suitable, vocabulary-wise, popular science books might be for use in *English for Science* classes or as an extra reading resource for this particular ESP field. Based on a 2-million-word corpus of popular science books we have compiled, we determine the vocabulary profile of this genre, including how many words are needed to reach both the minimum and the ideal reading comprehension levels, as well as how much high-frequency general vocabulary, academic vocabulary and technical vocabulary they contain. The results have pedagogical implications and recommendations are given regarding the needed proficiency level of the students for whom these texts might be of most use.

Key words: popular science books, English for Science, vocabulary profiling, word lists, reading comprehension

1. Introduction

Reading comprehension cannot be achieved without a certain level of vocabulary knowledge. Naturally, comprehension goes far beyond just knowing the meaning of the words used in a text, but such knowledge is normally considered a precondition for the understanding to take place. In this study, we focus on the genre of popular science books, i.e., the complexity and the level of their vocabulary, and try to assess their usability in the *English for Science* classes. *English for Science* is typically taught as an academic course in tertiary education of science and engineering departments. In this paper we will focus on the vocabulary needs of these students and try to assess how helpful the reading of science popular books may be for such students.

[*] Faculty of Philology, University of Montenegro, Danila Bojovića bb, 81402 Nikšić, Montenegro; e-mail: vmilica@ucg.ac.me

2. Theoretical background

We start this section by explaining the role of vocabulary knowledge for reading comprehension and then proceed with a brief review of some of the word lists which have been produced for the ESL and ESP purposes to date.

2.1 Knowledge of vocabulary and reading comprehension

One of the generally accepted assumptions is that vocabulary knowledge can predict the level of reading comprehension (Laufer and Ravenhorst-Kalovski 2010). In fact, the knowledge of vocabulary is the single best predictor of how much a reader will understand the text read (Naggy 1988). With this in mind, many researchers have tried to determine how many words a reader should know in order to be able to understand certain types of texts at certain levels.

One of the seminal studies in this respect is that of Laufer (1989), in which she determined that a 95%-vocabulary threshold would be needed to achieve a ‟reasonable“ level of reading comprehension (it is assumed that part of the remaining 5% of the words could be guessed from the context). An equally cited study is that of Nation (2006), in which the bar is raised higher – at 98%, which should ensure ‟optimal” reading of texts.

The thresholds quoted seem very high for those learning English as a second language. Nation (1990) argues that a 98%-threshold would typically assume knowledge of up to 9,000 word families (a word family includes the root word and words derived from it, e.g. *graduate, undergraduate, postgraduate, graduated, graduation* etc.). Unfortunately, for many such speakers these levels seem unattainable. However, some researchers have shown that the number of the words needed depends on the type of text and proved that a good selection of words for certain purposes may substantially reduce the word count needed.

2.2 Word lists

The idea behind word lists is to make selections of words according to their priority for certain purposes, where the priority is equated with their frequency in certain types of texts. In the short review below we will focus on just several such lists.

The most famous word list was developed in 1953 by West. It is called the *General Service List*, abbreviated to the GSL, and it contains about 2,000 word families which are judged as most frequent in general English. When learning English as a second language these are usually first covered and many ESL speakers will know most of them when they have reached a certain level (for instance, the intermediate level). In general texts, the GSL typically covers around 80% of all the words used (Nation 2001).

Since the GSL, many other general word lists have been made. Among them, it is worth mentioning the set of 25 word lists derived from the BNC and the COCA corpora by Nation (2012). These 25 word lists contain 1,000 word families each, sorted by their frequency in the said corpora, whereas the additional four lists accompanying the set include proper names, marginal words (swear words, exclamations, letters of the alphabet), transparent compounds (whose meaning can be easily guessed if one knows the meanings of their elements), and abbreviations.

An almost equally famous word list is the *Academic Word List* (the AWL), compiled by Coxhead in 2000. It contains 570 word families which are outside the GSL and are frequently found in various kinds of academic texts. Those learning *English for Academic Purposes* will typically focus on them, assuming they have first mastered the GSL words. In academic texts, the AWL typically covers around 10%, as various studies have shown (Coxhead 2000; Chen and Ge 2007; Vongpumivitch et al. 2009; Valipouri and Nassaji 2013, etc.). The 2,500 word families of the GSL and the AWL combined, cover around 90% of the words in a typical academic text. This seems as a good coverage, but it is still not sufficient for either reasonable or optimal reading comprehension (the thresholds for which are 95% and 98%, respectively).

Many variations of academic word lists have been made since, as have other lists been tailored for various specific purposes. Bearing in mind that in this paper we focus on the needs of the *English for Science* students, the most relevant word list for us in this case is the *Science List* (the SL), produced in 2007 by Coxhead and Hirsch. It contains 318 word families which are outside the GSL and the AWL, and are frequently found in various science texts (the corpus it was derived from includes the following disciplines: agricultural sciences, biology, chemistry, computer science, ecology, engineering & technology, geography, geology, horticultural science, mathematics, nursing & midwifery, physics, sport and health science, veterinary and animal sciences). Its coverage in these science texts was around 4% (Coxhead and Hirsch 2007). This means that fewer than 3,000 word families of the GSL, AWL and SL combined, would cover up to 94% of the words in a scientific academic text, which is close to what is needed for reasonable comprehension. Assuming that students have some prior knowledge of general vocabulary, the quoted word count does not sound unattainable, and it can be imagined that most of vocabulary contained in the three lists could be realistically covered in the *English for Science* classes.

3. Corpus and methodology

The popular science books corpus compiled for this study contains 15 books. The titles were chosen by popularity, as indexed by the *GoodReads* website (www.goodreads.com), the section of *popular science books* (as of January 2019). Choosing popularity as the inclusion criterion meant that not all disciplines would be equally represented and that some might be overrepresented; however, we still chose popularity as the inclusion criterion as popularity is often the sole criterion based on which readers choose this type of books, and we also assumed that the most popular books would also be most accessible to them. For many sciences there will not be many or almost any popular science books, as these are not likely to be best-sellers, so opting for the equal representation in this type of a corpus would not make much sense. Additionally, the *Science List* was also based on a selection of sciences, although a much more comprehensive one, as probably no science corpus can be comprehensive enough – much depends on how we define *science*, after all. The list of the books which are part of our corpus, together with their word counts, is given in Table 1 below:

Science popular book	**No. of tokens**
A Brief History of Time (by Stephen Hawking)	64,135
A Short History of Nearly Everything (by Bill Bryson)	173,487
The Selfish Gene (by Richard Dawkins)	163,313
Cosmos (by Karl Segan)	126,699
The Origin of Species (by Charles Darwin)	198,833
The Elegant Universe (by Brian Greene)	151,258
Astrophysics for People in a Hurry (by Neil Degrasse Tyson)	33,144
What If?: Serious Scientific Answers to Absurd Hypothetical Questions (by R. Munroe)	64,679
The Demon-Haunted World: Science as a Candle in the Dark (by Carl Segan)	146,919
Physics of the Impossible (by Michio Kaku)	107,280
A Universe from Nothing (by Lawrence Krauss)	275,589
The Trouble with Physics (by Lee Smolin)	143,344
Dreams of a Final Theory (by Steven Weinberg)	107,590
Just Six Numbers: The Deep Forces That Shape the Universe (by Martin Rees)	51,990
The Emperor's New Mind (by Roger Penrose)	195,455
TOTAL	**2,003,715**

Table 1. The popular science books corpus

In this study, we used *AntWordProfiler 1.4.0w* (Anthony 2014), a vocabulary profiling software which enables corpora to be compared against word lists.

The word lists used for comparison are those discussed in the introductory section.

4. Results

We first compared our corpus against the General Service List, the Academic Word List and the Science List. The results are presented in Table 2 below:

GENRE	**SCIENCE POPULAR BOOKS**	
Word list	TOKEN%	CUM %
GSL 1st 1000	79.07	79.07
GSL 2nd 1000	4.49	83.56
AWL	5.32	88.88
SL	1.51	90.39
Off-lists	9.61	100

Table 2. Lexical profile of the *Popular Science Book Corpus*

Table 2 above informs us that the most frequent general words make up 83.56% of the corpus, which is within the expected ranges (as discussed in the introduction). More than 4 in 5 words will be familiar to the reader who just knows the most frequent 2,000 words of English.

The frequency of academic words in the *Popular Science Book Corpus* is half that featured in academic texts, which is significantly lower. Still, this frequency is considerably higher than that which could be expected in a general text. This means that students will encounter more academic vocabulary in popular science books than in general texts, but they will not be able to substantially improve their academic vocabulary by only reading this type of books.

Technical science words make up just 1.51% of the vocabulary used in popular science books, which is not much. We assume that the authors of these books deliberately avoid technical vocabulary, which is why there is three times less of it in such books than in academic research papers and academic books, i.e., academic texts.

Off-lists contain the letters of the alphabet, symbols, proper names, abbreviations, transparent compounds and not so frequent words. The last group – the not so frequent words group, presumably takes the smallest proportion of the remaining 10% of the words, which means that the student who knows most of the general vocabulary, half of the most frequent academic words and a third of the most frequent technical words, would have no problem reading popular science books. If the aim for the students is to become familiar with only the most frequent academic and technical words, then supplementing the class and self-study materials with the materials extracted from such a source makes a lot of sense. However, for those students intending to really read academic texts (research papers and books), much more vocabulary would be needed. This means that relying on this type of materials as core readings for such students would be a mistake; in this case, at best, they could be occasional supplementary materials.

For illustration purposes, we will quote a short stretch from the corpus, highlighting the words in various ways depending on the word list they belong to:

> **A Moment of Truth**
>
> **We all met at the** Institute **Saturday morning as planned. It was a bright sunny morning, and the** *atmosphere* **was** jokingly relaxed. **I, for one, half expected that** Aspinwall **would not show up; once he did, I spent** 15 minutes extolling **the** import **of this first weekend he had come into the office. He** assured **me it wouldn't happen again.**
>
> **We all** huddled **around** Morrison**'s** computer **in the office he and I shared**. Aspinwall **told** Morrison **how to bring his** program **up on the** screen **and showed us the** precise **form for the** required input. Morrison appropriately formatted **the results we had** generated **the** previous **night, and we were set to go**.
>
> **The particular** calculation **we were** performing **amounts, roughly speaking, to determining the mass of a certain particle** *species* **– a** specific *vibrational* pattern **of a** string **– when moving through a** universe **whose** Calabi-Yau component **we had spent all** fall identifying. **We hoped, in line with the** strategy discussed **earlier, that this mass would agree** identically **with a** similar calculation **done on the** Calabi-Yau **shape** emerging **from the space-tearing** flop transition. **The latter was the relatively easy** calculation, **and we had completed it weeks before; the answer turned out to be** 3, **in the particular units we were using. Since we were now doing the** purported mirror calculation **numerically on a** computer, **we expected to get something** extremely **close to but not** exactly 3, **something like** 3.000001 or 2.999999, **with the** tiny **difference arising from** rounding errors.
>
> (From the *Elegant Universe*, by Brian Greene)

The words marked in bold, which overwhelmingly prevail in this stretch, belong to the first 1,000 words of the GSL; the double-underlined words belong to the second 1,000 GSL words. The single-underlined words, including such words as *institute, assure, generate, component, error*, *transition*, belong to the academic word list set. Only three words are marked in italics (*atmosphere, species, vibrational*) and these belong to technical, i.e. scientific words. The remaining words are off-lists; as can be seen, these include proper names (*Morrison, Aspinwall...*), numbers, and other words. In *English for Science* classes with at least intermediate students, teachers should focus on the words such as those single-underlined or in italics above.

In Table 3 we see how our corpus fares against another set of word lists – that of the 29-word-list set derived from the BNC and the COCA corpora (Nation 2012).

BNC lists	**POP BOOKS**	
	TOKEN%	CUM. %
Proper nouns	1.41	1.41
Marginal words and letters of the alphabet	9.01	10.42
Transparent compounds	0.21	10.63
Abbreviations	0.09	10.72
1st 1,000 word families	68.25	78.97
2nd 1,000 word families	7.71	76.68
3rd 1,000 word families	5.97	92.65
4th 1,000 word families	2.00	94.65
5th 1,000 word families	1.07	95.72
6th 1,000 word families	0.63	96.35
7th 1,000 word families	0.51	96.86
8th 1,000 word families	0.40	97.26
9th 1,000 word families	0.25	97.51
10th 1,000 word families	0.19	97.7
11th 1,000 word families	0.13	97.74
12th 1,000 word families	0.09	97.83
13th 1,000 word families	0.08	97.91
14th 1,000 word families	0.05	97.96
15th 1,000 word families	0.06	98.02
16th-25th 1,000 word families	1.98	100.00

Table 3. The level of vocabulary in *Popular Science Books Corpus*

From Table 3 we learn that proper names, marginal words with the letters from the alphabet, abbreviations, and transparent compounds make as much as one tenth of the corpus. These are typically assumed transparent to the reader, with no need of additional clarification.

Reasonable reading comprehension is achieved at a level between the first four or five thousand most frequent word families. A cross-comparison of the results with Table 2 shows that more than 90% of those words will be general words, and just over 8% of them would be academic and technical words. This shows us that popular science books are not ideally suited for use in the *English for Academic Purposes* and *English for Science* classes. Even though *English for Science* teachers may be tempted to use such materials, as these are interesting and much more accessible to the teacher if he/she does not have much expertise in the scientific field concerned (which is often the case), these materials will be just slightly better than general texts of any kind and, in fact, of little use to those students truly pursuing the learning of English for the purpose of reading academic papers and books.

Ideal reading comprehension is achieved at a high level – the knowledge of the 15,000 most frequent words, which means that only truly proficient ESL readers can ideally read these books. Ideal comprehension level is therefore not the target one can realistically pursue in the limited number of hours devoted to the *English for Science* classes.

4. Conclusion

In this paper, we discussed the value of popular science books, vocabulary-wise, for use in the *English for Science* classes. We determined the lexical frequency profile of our popular science books corpus against several word lists and established that popular science books contain a significant amount of general vocabulary, but also half of the most frequent academic and a third of most frequent technical (scientific) words. We conclude that these texts have little value when it comes learning the vocabulary needed for reading real science and may not be of use as core texts or on a frequent basis in the *English for Science* classes. They may, however, be used to occasionally supplement the materials, as a source of general words and some of the basic academic and technical words, which is perhaps best suited for those students just entering the *English for Science* courses, i.e. transitioning from the *English for General* to *English for Specific Purposes*.

References

Anthony, L. (2014). *AntWordProfiler* (Version 1.4.1) [computer software]. Tokyo: Waseda University. (8 August 2017) <http://www.laurenceanthony.net/software>.

Chen, Q. and G. Ge (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles. *English for Specific Purposes*, 26, 502–514.

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

Coxhead, A. and D. Hirsh (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliquée*, 12(2), 65–78.

Laufer, B. and G. C. Ravenhorst-Kalovski (2010). Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a foreign language*, 22(1), 15–30.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension. In: C. Lauren and M. Nordman (eds.), *Special language: From humans thinking to thinking machines*, Clevedon: Multilingual Matters, 316–323.

Nagy, W. (1988). *Teaching vocabulary to improve reading comprehension*. Newark: International Reading Association.

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge: Cambridge University Press.

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

Nation, I. S. P. (2012). *The BNC/COCA word family lists*. (1 April 2018) <http://www.victoria.ac.nz/lals/about/staff/paul-nation>.

Valipouri, L. and H. Nassaji (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes*, 12(4), 248–263.

Vongpumivitch, V., J. Huang and Y. Chang (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, 28, 33–41.

West, M. P. (ed.) (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longmans: Green.