REGISTER JOURNAL – Vol 17, No 1 (2024), pp. 164-188
DOI: http://dx.doi.org/10.18326/rgt.v17i1.164-188
p-ISSN: 1979-8903; e-ISSN: 2503-040X

A Corpus-Based Lexical Coverage of Course books in Nigeria: A Case Study

Hamisu Hamisu Haruna[1]*, Azza Jauhar Ahmad Tajuddin[2], Ibrahim Bashir[3]

[1]Department of English, College of Humanities, Al-Qalam University Katsina - Nigeria
[2]English Language Learning Center, Center of Fundamental and Continuing Education, Universiti Malaysia Terengganu – Malaysia,
[3]Department of English, Northern College of Nursing, Arar – Saudi Arabia

*) Corresponding Author
Email: azzajauhar@umt.edu.my
DOI: 10.18326/rgt.v17i1.164-188

Submission Track:
Received: 09-01-2024
Final Revision: 09-05-2024
Available Online: 06-06-2024

Abstract

Since vocabulary is one of the most important components of reading comprehension, the relationship between the two has been studied in great detail. The significance of this relationship lies in the degree of coverage of the word families in the texts. This study examined the lexical coverage of a corpus of 6,802,300 words from the first-year course books of the National Open University of Nigeria. With Anthony's AntWordProfiler software, we analyzed the lexical coverage of the corpus using the Lexical Frequency Profiling approach. The current study used Nation's (2012) BNC/COCA to determine the necessary vocabulary size for course book comprehension. The corpus study revealed that in order to

reach 95% and 98% of the entire course book corpus, respectively, 5000 and 11000 word-families were required. However, vocabulary size needed for comprehension of each disciplinary field varied greatly, with the hard sciences having a significantly higher lexical demand as compared to the other fields. Students need a larger vocabulary to interact with and understand the course books, especially in the hard sciences. Therefore, materials writers and instructors should consider the specific disciplinary vocabulary needs in course books. Similarly, due to disciplinary differences, more specific instructions and glossaries are needed for first-year university students better to understand course books, especially hard science course books. The study demonstrated the significance of corpus-based approaches in the analysis of language learning materials. Overall, the study underlined the importance of sufficient vocabulary for reading comprehension.

Keywords: vocabulary, lexical coverage, lexical profile, lexical load, course books

INTRODUCTION

Recently, corpus-based research has focused on the analysis of academic writing, which has led to a growing number of students having to read course books in English. This has resulted in the identification of particular features of academic vocabulary in course books, which have been of great concern in numerous EAP studies (Valipouri & Nassaji, 2013). This is because English course books, especially in ESL/EFL classes where they effectively function as a syllabus, are crucial for students' vocabulary acquisition (Yang & Coxhead, 2022).

Conversely, vocabulary knowledge breadth and depth—is critical to learners' performance in the four skills (Janebi Enayat & Derakhshan, 2021; Miralpeix & Muñoz, 2018). Many studies have demonstrated the strong correlation between successful reading comprehension and vocabulary knowledge, especially breadth (Zhang & Zhang, 2020). Indeed, there is evidence that a large receptive vocabulary one facilitates good reading comprehension recognises regardless of mastery level (Nation, 2006). Research has shown that vocabulary size and reading comprehension positively correlate (Laufer & Ravenhorst-Kalovski, 2010).

Learners in an ESL situation, then, need to be accustomed to this knowledge in order to acquire vocabulary (Haruna, et al., 2018; Ibrahim et al., 2018). In many contexts, this amount of input—especially in language learning texts—could be the primary means of vocabulary growth, especially if it is relatively low (Sun & Dang, 2020). In order to maximize learning and enable readers to both understand the content of the text and focus on the words that are most useful to them, the vocabulary used in these writings should be carefully selected. For this reason, much research has looked at vocabulary in ESL/EFL course books.

Nevertheless, there are still research trends on vocabulary that are important for the current study especially in the context of course books. Lexical load, academic vocabulary, and frequency are some of these concepts. Sun and Dang (2020), for example, distinguish three different approaches to vocabulary analysis in their study of academic writing. Considering the positive relationship between lexical knowledge and comprehension in research, the study of vocabulary in academic texts based on lexical coverage is believed to be significant as it provides information on how students can understand textbooks (van Zeeland & Schmitt, 2013). This is true even when learners' comprehension increases as their vocabulary knowledge increases (Schmitt, Jiang & Grabe, 2011).

The second trend in studies focuses on the calculation of high-frequency words used in texts. This type of research is crucial because it examines how many words are good for students. For example, research found that only 1400 of the 2000 word families were included in the course book series. In the same vein, it was reported that the New English File textbooks contained only 1435 of the 2000 word families (Su & Dang, 2020). The third area of study is repetition in text materials. Research on this is essential because vocabulary acquisition benefits from repetition (Webb & Nation, 2017).

Despite the abundance of studies on vocabulary in ESL contexts, there is dearth of studies on lexical coverage on course books especially for open universities developed for self-learning.

Since English language is the vehicle through which all forms of education are imparted and all forms of formal education in Nigeria are acquired, it continues to be the common language of all ethnic groups despite individual and cultural differences. Vocabulary mastery is therefore essential especially for Open University students who have physical limitation and largely depend on course books for their academic programmes. The aim of this study, therefore, is to find out the vocabulary that students in the National Open University of Nigeria are exposed to by examining the vocabulary coverage of first-year course books. In this study, first-year course books were chosen because they are the first materials that students come into contact with during their first year programme after their transition from secondary school to university (Makino, 2024). The results are relevant to educators, course developers and eventually students studying at undergraduate level. The next section reviews the related literature.

Lexical Coverage

The term "lexical coverage" is defined as the reader's familiarity with the running words of the text (Nation, 2006, p.61). Lexical coverage is the extent to which the target audience is familiar with the input words (Webb, 2021). It measures the degree of ease or adequacy with which a text can be understood and assesses the extent to which a text has been learned or understood (Hsu, 2014). Therefore, the question of how many words in a text need to be understood in order to achieve adequate or reasonable comprehension, thereby avoiding the need for learners to use "compensatory strategies" (Laufer, 2013, p. 868), and to enable incidental vocabulary learning is a common way in which vocabulary research addresses the issue of lexical coverage (Nation, 2006).

Extensive research has shown the importance of lexical coverage for understanding written texts. Furthermore, research shows that students' reading comprehension improves when they recognise more terms in a book (Arndt, 2022; Laufer & Ravenhorst-Kalovski, 2010; Schmitt, Jiang & Grabe, 2011). The influence that vocabulary knowledge has on comprehension is therefore

highlighted by studies on lexical coverage (Webb, 2021). Lexical coverage generally corresponds to the degree of familiarity readers have with the words occurring in the text. It indicates whether or not a text has been sufficiently learned and understood and whether it can be easily or sufficiently digested (Laufer & Nation, 1995).

Research on lexical profiling therefore examines the vocabulary knowledge required to obtain 95% and 98% for comprehension (Laufer & Ravenhorst-Kalovski, 2010). In response to these studies, research (e.g., Dang & Webb, 2014) investigated the number of words required for comprehension of different discourse styles. For example, Nation (2006) found that spoken discourse requires 6,000–7,000 word families to achieve 98% lexical coverage, while written literature requires 8,000–9,000 word families.

In light of the above postulates, positive relationship were demonstrated on lexical depth and reading comprehension in many studies (e.g. Laufer, 2013; Nurmukhamedov & Webb, 2019; van Zeeland & Schmitt, 2013), which has led to wide-ranging considerations of lexical profiling and coverage. According to Laufer (1995), adequate comprehension is defined as a vocabulary comprehension of 95% minimally. However, with less than 95%, comprehension is likely to be inadequate (Coxhead & Boutorwick, 2018). This means that learners with a vocabulary of 95% or less may need some support when reading a text. In contrast, learners who can read a text independently are those who have mastered 98% of the vocabulary. This is because knowledge of vocabulary is considered the key to understanding a written text (Rabadi, 2023).

Consequently, numerous studies have investigated lexical coverage in textbooks using Nation's BNC/COCA frequency lists. Nguyen (2020) examined the lexical components in 30 units of English textbooks of Vietnamese high school students using Cobb's Lextutor and the 25 1000-word lists from Nation's BNC/COCA frequency lists (Nation, 2016). According to the results, students needed knowledge of 3,000-5,000 word families to understand 95% and 98% of the textbooks, respectively, which was above their vocabulary knowledge of 2000 high-frequency words. In addition,

the textbooks did not sufficiently expose students to new words in context.

In another study, Sun and Dang (2020) analyzed the lexical load of a collection of Chinese high school textbooks produced by Yilin Press. These textbooks were used by 265 high school students. Based on the nation's BNC/COCA lists, Sun and Dang's (2020) used the Updated Vocabulary Levels Test (UVLT) (Webb et al., 2017). According to the study, students needed 3,000 and 9,000 word families to fully understand 95% and 98% in the texts. Based on the test takers' responses in the vocabulary test, this vocabulary load was significantly higher than their level of vocabulary comprehension. It was also found that of the 265 students, only five had learned the 3,000 word families and the majority had difficulties mastering them at their level.

To summarize, while the above review has shown that vocabulary knowledge is a strong predictor of reading comprehension, most of these efforts have focused on improving the accessibility of textbooks produced by foreign publishers for use by second language learners around the world (such as: Sun and Dang, 2020; Coxhead & Boutorwick, 2018). Therefore, it is still unclear how much high-, mid- and low-frequency vocabulary in locally published course books contribute to students' overall reading comprehension, particularly in African context. This indicates that vocabulary research in Nigerian university course books is under-researched. Thus, to the researchers' knowledge, there are no studies on lexical coverage in the course books of National Open University. Against this background, the present study attempts to fill in the gap.

In this light, this study makes important contributions to the existing body of knowledge through an in-depth examination of lexical coverage in the course books of National Open University of Nigeria (NOUN). Findings provide valuable insights with practical implications that enhance our understanding of language for academic purposes in open universities and offer suggestions for improving the educational experience for Nigerian students. The implications of this research go beyond the immediate context of

NOUN and enrich the broader discourse on vocabulary research and pedagogy in open education. The next section discusses the lexical frequency profile in relation to the study.

Lexical Frequency Profile

Lexical Frequency Profiling (LFP) is an analytical method created in 1995 by Laufer and Nation that measures both the lexical depth of a text and the useful size of a learner's vocabulary (Stamatović, Bratić & Lakić 2020). LFP is widely used in EFL as well as ESL research and education, often to ascertain the lexical density of particular texts, as it is the most well-known frequency-based measure of vocabulary analysis. It has been discovered that the LFP, while not the only way for measuring lexical richness, yields results that are very similar compared to alternate methods (Lindqvist, Anna & Camilla, 2013).

In contrast to lexical coverage studies, lexical profiling, which is defined as the estimate number of vocabulary required for 95% and 98% word coverage (Webb, 2021) and mastery of the most common terms in English, has been the subject of a significant amount of research (Nation, 2006). The lexical load of numerous discourse genres has been the subject of such investigations. Studies on the vocabulary demands of EFL textbooks (Sun & Dang, 2020), the lexical load of television shows and films (Webb & Rodgers, 2009), the vocabulary profile of spoken academic English (Dang & Webb, 2014), and even the vocabulary profile of popular songs in English Language Teaching (ELT) (Tegge, 2017) have all been studied. These studies provide clarity on the number of words needed to understand different types of speech and advise teachers on what vocabulary to teach in terms of frequency ranges.

The present aim of the study, therefore, is to analyze the lexical coverage of the National Open University of Nigeria's first-year course books using Nation's (2012) BNC/COCA word list. Nation's (2012) BNC/COCA word list is adopted as it is regarded as an extensive reference corpus that has been used for many years (Hsu, 2013). As far as we know, such a study has never been conducted in the Nigerian context. Based on the findings of this study, a vocabulary learning target will be set for undergraduate university

students, in line with Webb and Nation (2012) who recognise the value of conducting lexical profile studies to determine vocabulary targets for learning education.

This study aims to address following research questions:

RQ1: What are the lexical coverage and Schmitt and Schmitt's (2014) coverage of high, mid and low frequency in first-year university course books?

RQ2: What are the variations in lexical coverage across hard sciences, soft sciences, and non-sciences course books?

RESEARCH METHOD

The Compilation of the Corpus

For this study, a course book corpus of 6,802,300 words was compiled from three disciplinary fields developed by the National Open University of Nigeria (Hard Sciences, Soft Sciences and Non-Science). Through the website NOUN e-courseware (https://nou.edu.ng/e-courseware/), we obtained pdf versions of all the course books used in the corpus. All texts were initially contained in PDF files and were first converted to txt files using Anthony's (2022) AntFileConverter software. Once the files were converted, the data was manually cleaned, i.e. prefaces, tables, references, captions and appendices were removed before being analyzed using the AntwordProfiler tool. The cleaning of the texts was a necessary step before conducting the analysis in order to standardize the corpus (Benson & Coxhead, 2022; Chen & Ge, 2007; Lu & Coxhead, 2020).

Table 1. Word Counts in NOUN Course books Corpus

No.	Fields	Number of Books	Number of Words
1	Hard Sciences	75	2,370,502
2	Soft Sciences	50	2,030,393
3	Non-Sciences	75	2,401,405
	Total	200	6,802,300

The base word lists of the British National Corpus and Nation's Corpus of Contemporary American English (BNC/COCA) (2012) were used with the tool to analyze the corpus using word family as a unit of counting. Nation's (2012) BNC/COCA base word lists were used due to their size and ability to analyze vocabulary at multiple frequency levels. The corpus was created taking into account its representativeness, specificity, use of full texts, and electronic availability, as recommended by Sinclair (1991) and Barnbrook (1996).

Data Analysis

To determine the coverage of the NOUN corpus for the first year, we used the Nation's BNC/COCA 25,000 base level wordlists along with supplementary lists (proper nouns, marginal words, compound words and abbreviations). The adoption of the BNC/COCA base word lists was motivated by the desire to create a classification of word frequency from most frequent to least frequent. It is used to assess the lexical threshold/load required to understand certain types of text within a given discourse domain. Prior to the final analysis, words that were not included in Nation's (2012) base lists or the supplementary lists were reclassified into the appropriate lists. Examples of such words are proper nouns (e.g. Nigeria, Yoruba) and compound words (e.g. antenatal, tapewarm).

The corpus was then analyzed using Anthony's (2021) AntwordProfiler software. This software allows users to input written material and analyze its lexical coverage based on specific base word lists. The software also displays the frequency and scope of each word based on word lists. The result for the coverage for each 1000-word level was added until the thresholds of 95% and 98% were reached without supplementary lists. On the other hand, coverage of each 1000-word level was also added to supplementary lists until the overall coverage reached the thresholds of 95% and 98% respectively. Cumulative coverage included proper names, marginal words, compound words, and abbreviations because previous studies (e.g., Dang, 2019; Sun, & Dang, 2020) have found that they require little learning. This is followed by the coverage of

Schmitt and Schmitt's (2014) high, mid, and low frequency vocabulary. Then the variations of lexical load in three fields (hard, soft and non-sciences) are examined.

RESULTS & DISCUSSION

Lexical Coverage and Schmitt and Schmitt's (2014) coverage of high, mid and low frequency

To answer this, Table 2 presents the lexical profile in accordance with BNC/COCA base word lists, showing the number of families, coverage and the three most frequently occurring words in each. The distribution of the vocabulary of the NOUNC corpus among the 25 BNC/COCA base word lists and the four supplementary lists, adopting the 25 BNC/COCA base word lists, is shown in this table.

Table 2. Lexical Profile of NOUNC across Nation's (2012) BNC/COCA Word Lists

Base word lists	Running words	Coverage %	Cumulative Coverage %	Word Family	Examples
1st 1000	4619909	67.92	67.92	1000	The, of, and
2nd 1000	848946	12.48	80.4	999	Unit, development, social
3rd 1000	648340	9.53	89.93	1000	Objectives, data, assessment
4th 1000	144201	2.12	92.05	998	tutor, acid, acids
5th 1000	77079	1.13	93.18	993	membrane, nutrients, enzymes
6th 1000	50094	0.74	93.92	975	Pre, logistics, hygiene
7th 1000	34789	0.51	94.43	946	Amin, glucose, microbial
8th 1000	26214	0.39	94.82	912	Skeletal, sanitation, enumerate
9th 1000	17726	0.26	95.08	845	microorganisms, lipids, germination
10th 1000	13165	0.19	95.27	787	Inter, respiration, iodine
11th 1000	11596	0.17	95.44	723	Anti, cheque, yam
12th 1000	8800	0.13	95.57	644	Cytoplasm, phylum, protozoa
13th 1000	7566	0.11	95.68	593	Versa, cassava, viz
14th 1000	7637	0.11	95.79	519	Intramural, sharia, alveolar
15th 1000	5024	0.07	95.86	479	Hausa, organelles, tillage

Base word lists	Running words	Coverage %	Cumulative Coverage %	Word Family	Examples
16th 1000	5137	0.08	95.94	453	Neo, quran, covalent
17th 1000	4377	0.06	96	395	Prokaryotes, naira, neutrons
18th 1000	3457	0.05	96.05	358	Flagella, endoplasmic, unicellular
19th 1000	2603	0.04	96.09	327	Malware, alkyl, redox
20th 1000	2041	0.03	96.12	305	Gametophyte, hydrides, vires
21st 1000	2518	0.04	96.16	251	Socio, xylem, endosperm
22nd 1000	1285	0.02	96.18	222	Meristem, alkane, paramecium
23rd 1000	930	0.01	96.19	199	Ethno, auxin, auxins
24th 1000	1010	0.01	96.2	165	Burette, alkynes, hypercholesterolemia
25th 1000	1263	0.02	96.22	170	phloem, alkene, monosaccharides
proper nouns	96903	1.42	97.64	5164	Nigeria, Africa, English
Marginal words	41893	0.62	98.26	39	Mm, xo, e
Transparent compounds	18263	0.27	98.53	1126	feedback, healthcare, classroom
Abbreviations	18034	0.27	98.8	634	ict, etc, uk
Not Found in the Lists	81500	1.2	100	33125	Prokaryotic, anansewa, ilos
Total	6,802,30			55,346	

The table shows that lexical coverage decreases through the first 1,000 to 25,000 word families, with 67.92%, 12.48% in the first and second 1000 high frequency words. However, the coverage rate dropped significantly to 2.12% and 1.13% for the fourth and fifth 1000-word family base word lists, which nevertheless contained a reasonable number of word families. After the fifth 1,000-word list, the number of participating word families decreased noticeably and the coverage rate increased (less than 1%) only slightly until the coverage rate of the Nation's (2012) BNC/COCA lists and the supplementary lists combined reached 98.8%.

Additionally, the four supplementary lists of the Nation (2012) BNC/COCA lists provided insightful information about the NOUNC corpus. A percentage of 1.42% of the course books contained proper

nouns. These were given to people and places, such as *Nigeria* and *Lekki*. In addition, there were other lists that included coverage of abbreviations in the corpora (0.27%), marginal words (0.62%) and transparent compound words (0.27%).

On the other hand, Table 3 displays the lexical coverage of the NOUNC corpus. Taking into account the modified supplementary lists from Nation (2012), the vocabulary analysis opted for Laufer & Ravenhorst-Kalovski (2010) 95% and 98% cut-off point. These were made part of the analysis because, as Nation (2013) notes, students have no difficulty learning supplementary lists once they have mastered them (see, e.g., Dang and Webb 2014; Nation, 2006; Coxhead, Dang& Mukai, 2017). Because of their high coverage, the words in the supplementary lists were critical to achieving 95% and 98% coverage.

Table 3. Cumulative Coverage of the NOUNC by BNC/COCA Base lists

Base word lists	Coverage (%)	Cumulative Coverage with supplementary lists
proper nouns	1.42	1.42
Transparent compounds	0.27	1.69
Abbreviations	0.27	1.96
marginal words		2.58
	0.62	
1st 1000	67.92	70.5%
2nd 1000	12.48	82.98%
3rd 1000	9.53	92.51%
4th 1000	2.12	94.63%
5th 1000	1.13	**95.76%**
6th 1000	0.74	96.5%
7th 1000	0.51	97.01%
8th 1000	0.39	97.4%
9th 1000	0.26	97.66%
10th 1000	0.19	97.85%
11th 1000	0.17	**98.02%**

As shown in Table 3, students needed to have a vocabulary of 5,000 most frequently occurring word families from the BNC/COCA base lists plus the supplementary lists to attain the 95% minimum comprehension. This result is in line with Hsu's (2014) who found that engineering textbooks had coverage of 95% in vocabulary analysis at 5,000 words. In the same vein, a vocabulary of 11,000 most frequently occurring word families from the

BNC/COCA base lists and the supplementary lists were needed to attain 98%, which is considered the optimal vocabulary for reading comprehension in the corpus. This result is confirmed by previous research (e.g., Ng et al., 2020; Sun & Dang, 2020), which also achieved 98% on 11,000 word base lists.

In this regard, the table underpins the relevance of supplementary lists for comprehension which is demonstrated by the fact that the 98% coverage of the 25,000 base words would not be achieved without them. This claim was supported by a related discovery by Yang and Coxhead (2020), who also found that the third year book in the Chinese English textbook series achieved only 98% coverage on the 25,000-word base list without supplementary lists. Although the study was conducted in an EFL setting, such a conclusion may be extended to ESL context. This shows how important the supplementary lists are for understanding the corpus. In addition, the results show how lexically demanding the NOUNC corpus was, suggesting that students need a larger vocabulary to comprehend the course books.

When evaluating how well high, mid, and low-frequency words were covered in the NOUNC corpus, it is imperative to bear in mind their definitions. Schmitt and Schmitt (2014) contested Nation's base list classification, classifying the first three 1,000-word families (BASEWRD 1000-3000) of the BNC/COCA lists as high-frequency vocabulary, (BASEWRD 4000-8000) as mid-frequency vocabulary and (BASEWRD 9000–25000) as low-frequency vocabulary using Nation's frequency-based base word lists (2012). Next, the coverage of the BNC/COCA base word lists that were included in each band was added to determine the coverage of each band. For example, high-frequency words were estimated using the BASEWRD 1, 2, and 3 coverage, and the same method was used to compute the coverage of mid- and low-frequency vocabulary. As indicated in Table 4, this approach allowed researchers to obtain a more thorough grasp of the vocabulary distribution in the NOUNC corpus.

Table 4. Coverage of high, mid, and low-frequency

Frequency bands	Base word lists	Coverage
High-frequency vocabulary (1,000-3,000)	1, 2, 3	89.93%

Mid-frequency vocabulary (4,000-8,000)	4-8	4.89%
Low-frequency vocabulary (9,000-25,000)	9-25	1.4%
Proper nouns, marginal words, compounds, abbreviations	31-34	2.55%
Off-List	35	1.23%
Total		100%

Based on Schmitt and Schmitt (2014) classification, high frequency words covered the highest percentage of 89.93% of all tokens in the corpus' total coverage, as shown in Table 4 above. Low-frequency vocabulary covered the least percentage of the corpus, with 1.4%, whereas mid-frequency words made up 4.89%, a significantly lower but still considerable component of the overall coverage. The four supplementary lists, on the other hand, accounted for 2.55% of the running terms in the corpus.

Overall, the results of the lexical coverage and Schmitt and Schmitt (2014) coverage of high, mid and low frequency revealed three main issues. First, familiarity with the first, second, and third 1,000 high frequency word families is necessary to comprehend the NOUNC corpus. This finding aligns with recent research demonstrating the significance of the first 3,000 word families in spoken (Dang, Coxhead & Webb, 2017) and written discourse (Rugby, 2020). Second, in reaching 95% coverage in the NOUNC corpus, understanding supplementary lists was essential. Finally, the supplementary base word lists had more than 1% of words, indicating that certain fields may have technical words that are only present in that field, as was the case with finance according to Ha and Hyland (2017). Thus, for effective academic performance in ESL contexts, learners need an extensive vocabulary to meet the challenges of English in academic contexts for better reading comprehension (Szudarski, 2018). Variations of lexical load in hard sciences, soft sciences, and non-sciences are shown in the next section.

Variations in Lexical Coverage across Hard Sciences, Soft Sciences, and Non-Sciences Course books

To respond to this, different levels of coverage are required in each of the three areas to reach 95% and 98% of the NOUNC corpus. By comparing the vocabulary level of the course book sub-corpora with the Nation's base lists and determining the number of word families needed to reach 95% and 98%, the lexical text coverage of each field was determined. Words with little or no learning were also listed in Table 3, along with coverage of the base lists from the BNC/COCA corpus. These included swear words and exclamations, proper nouns (which were usually easy to identify), abbreviations (which were usually defined when they were first used in a text) and marginal words (which were only words consisting of letters of the alphabet).

The results revealed that knowledge of more than 4,000 supplementary lists was required to achieve a minimum criterion of 95% coverage for both soft sciences and non-sciences. This result is consistent with previous research studies (Dang and Webb, 2014; Dang, 2018a; Dang, 2018b) which showed that 95% was achieved for 4000 families. They also discovered that it takes a vocabulary of 3,000–4,000 word families to cover 95% of the soft sciences. However, they did not reach the same conclusions as Bratić & Stamatović (2021), who claimed a larger coverage of 9,000 families.

However, 8,000 plus supplementary lists would be required to achieve minimal coverage for the hard sciences, covering only 95.55% of the texts. This result suggests that the minimum vocabulary for reading comprehension in the hard sciences must be 8,000 words. Bratić & Stamatović (2021) also observed this high lexical demand. They found that reading a scientific, technical, or medical corpus required more than the 25,000 basic lists plus proper nouns, abbreviations, and marginal words.

Table 5. Lexical Coverage of First-year NOUN course books Across the three fields

BNC/COCA base word	Hard Sciences %	Cum %	Soft Science %	Cum %	Non-Sciences %	Cum %
proper nouns	0.8	0.8	1.69	1.69	1.82	1.82
Marginal words	1.01	1.81	0.33	2.02	0.47	2.29
Transparent compounds	0.35	2.16	0.23	2.25	0.22	2.51
Acronyms	0.37	2.53	0.2	2.45	0.21	2.72
1st 1000	64.38	66.91	68.26	70.71	71.12	73.84
2nd 1000	12.48	79.39	13.54	84.25	11.59	85.43
3rd 1000	9.72	89.11	10.6	94.85	8.45	93.88
4th 1000	2.58	91.69	1.81	**96.66**	1.92	**95.8**
5th 1000	1.63	93.32	0.86	97.52	0.88	96.68
6th 1000	0.98	94.3	0.56	**98.08**	0.65	97.33
7th 1000	0.68	94.98	0.35	98.43	0.48	97.81
8th 1000	0.57	**95.55**	0.25	98.68	0.31	**98.12**
9th 1000	0.43	95.98	0.14	98.82	0.2	98.32
10th 1000	0.3	96.28	0.11	98.93	0.15	98.47
11th 1000	0.27	96.55	0.1	99.03	0.13	98.6
12th 1000	0.23	96.78	0.05	99.08	0.09	98.69
13th 1000	0.21	96.99	0.05	99.13	0.07	98.76
14th 1000	0.2	97.19	0.04	99.17	0.08	98.84
15th 1000	0.15	97.34	0.02	99.19	0.04	98.88
16th 1000	0.16	97.5	0.03	99.22	0.04	98.92
17th 1000	0.13	97.63	0.02	99.24	0.04	98.96
18th 1000	0.11	97.74	0.01	99.25	0.02	98.98
19th 1000	0.09	97.83	0.01	99.26	0.01	98.99
20th 1000	0.06	97.89	0.01	99.27	0.02	99.01
21st 1000	0.08	97.97	0.01	99.28	0.02	99.03
22nd 1000	0.04	**98.01**	0	99.28	0.01	99.04
23rd 1000	0.03	98.04	0	99.28	0.01	99.05
24th 1000	0.04	98.08	0	99.28	0	99.05
25th 1000	0.05	98.13	0	99.28	0.01	99.06
Off lists	1.88	100	0.69	100	0.96	100

Conversely, the vocabulary required to cover 98% of the NOUNC corpus varied from soft sciences, which had the lowest lexical demands with 6,000 word families, to hard sciences, which had the highest demands with 22,000 word families. The results suggest that it may be more difficult for students majoring in hard sciences to reach the highest level of reading comprehension than for students majoring in other subjects (e.g. soft sciences and non-science). This result was supported by Vuković-Stamatovi (2020) and Bratic & Stamatovi (2021), who also found high lexical demand in hard sciences in their corpus. The former group felt that these

books were best read by truly proficient ESL readers. On the other hand, the latter group posits that the books can only be reserved for highly proficient graduates and professionals. The pedagogical implications are discussed in the next section.

Pedagogical Implications of the Main Findings

Based on the results presented above, the present study has some pedagogical implications. By comparing the lexical profile of the first-year course books NOUN with the modified BNC/COCA base word lists, this study was able to gain valuable insight into the nature of vocabulary in the corpus. In general, the study highlighted four important findings that could impact on the way first-year university students learn vocabulary for NOUN course books. First was the lexical profile of the corpus in terms of Schmitt and Schmitt's (2014) high, mid, and low frequency and Nation's (2012) BNC/COCA base lists. The results showed that 89.93% of the vocabulary in both corpora was high-frequency.

Second, the results showed that the discourse in hard science was significantly more lexically demanding than the discourses in the other two fields, which resulted in a considerable learning effort in vocabulary. Therefore, understanding the 22,000 word families and Nation's (2012) supplementary lists were necessary for 98% cumulative coverage in the hard sciences. In contrast, only 8,000 words were required in the non-sciences and 6,000 words in the soft sciences. Thus, because of the discrepancy between the three vocabularies, more specific instructions and glossaries were needed for first-year university students to understand the subject, especially in the hard sciences textbooks.

Third, as noted in other studies (e.g. Yang & Coxhead, 2020), course book resources have not been developed in line with research on how to improve students' vocabulary learning. Studies on vocabulary have provided suggestions and ideas for second or foreign language teaching (e.g. Nation, 2011), but their influence on course book production seems to have been small. It is also recommended that course book developers be made aware of and trained in tools such as online vocabulary profiles and that more

emphasis is needed on vocabulary in explicit instructions for publication.

Finally, the research has demonstrated the value of corpus-based approaches in the analysis of language learning materials. Because these methods shed light on text features, studies of this kind open up new perspectives for language training. Therefore, further corpus-based evaluations of educational resources are suggested. These analyses can, for example, examine the progression of learning across different learning stages or materials used in other educational settings.

CONCLUSION

Using Nation's (2012) BNC/COCA, this study examined lexical coverage, Schmitt and Schmitt's (2014) coverage of high, mid and low frequency as well as variations in lexical loads across hard sciences, soft sciences, and non-sciences in first-year university course books. To avoid being overwhelmed with level course books, students need to know the vocabulary requirements for each level. Course book developers can employ this analysis to methodologically control which words are included in their textbooks and which are omitted. For this study, we set a threshold of 95% text comprehension as the minimum for text coverage. The study found that the hard science textbooks had the highest vocabulary coverage at 22,000 words. As this is a stepping stone for the newly admitted student in the field, this is not surprising, and as a result, discipline-specific vocabularies are introduced.

To achieve the required vocabulary, students need to invest more time and effort in vocabulary learning to reach 95%. It might be easier to acquire the required vocabulary with the help of a manual or a list of words, which would also help in visualizing language acquisition. For this reason, ESL teachers in universities should provide vocabulary acquisition resources to first-year students so that they can cope with the enormous lexical load required for good reading comprehension. The results of this study show that there are a reasonable number of off-lists (1.2%). In view

of this, we suggest that future studies focus on how jargon is used in health science textbooks.

While this study provided valuable insights, it was limited to only one type of text: course books. This limitation restricts the usefulness and implications of the study. It is therefore recommended that future studies should examine lexical coverage in other types of texts, such as final year research projects and students' essays.

Acknowledgments

The authors would like to thank Universiti Malaysia Terengganu and Ministry of Higher Education, Malaysia for funding this project (FRGS/1/2020/SSI0/UMT/03/3) under the Fundamental Research Grant Scheme (FRGS) (vote number 59651).

REFERENCES

Anthony, L. (2021). AntWordProfiler (Version 1.5.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

Arndt, R. (2022). Vocabulary in digital science resources for middle school learners. *Applied Corpus Linguistics*, *2*(3), 100023. https://doi.org/10.1016/j.acorp.2022.100023

Barnbrook, G. (1996). *Language and computers: a practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Benson, S. (2020). Dot the pill down: investigating the linguistic needs of foreign rugby players and lexicon of spoken rugby discourse. *PhD Thesis*, Victoria University of Wellington.

Benson, S., & Coxhead, A. (2022). Technical single and multiword unit vocabulary in spoken rugby discourse. *English for Specific Purposes*, *66*, 111-130. https://doi.org/10.1016/j.esp.2022.02.001

Bratić, V., & Stamatović, M. V. (2021). Lexical profile of literary academic articles. *Ibérica*, (42), 115-138. https://doi.org/10.17398/2340-2784.42.115

Chen, Q., & Ge, G.-C. (2007). A corpus-based lexical study on frequency and distribution of coxhead's AWL word families in medical research articles (ras). *English for Specific Purposes*, *26*(4), 502–514.

Coxhead A (2017) Academic vocabulary in teacher talk: challenges and opportunities for pedagogy. *Oslo Studies in Language,* 9, 29–44.

Coxhead, A., Dang, T. N. Y., & Mukai, S. (2017). Single and multi-word unit vocabulary in university tutorials and laboratories: evidence from corpora and textbooks. *Journal of English for Academic Purposes*, *30*, 66-78.

Coxhead, A., & Boutorwick, T. J. (2018). Longitudinal vocabulary development in an EMI international school context: learners and texts in EAP, Maths, and Science. *TESOL Quarterly*, *52*(3), 588-610.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). Editors' introduction: conventions, terminology and an overview of the book. In H. Daller, J. Milton, & J. TreffersDaller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 1–32). Cambridge: Cambridge University Press.

Dang, T.N.Y., Webb, S., (2014). The lexical profile of academic spoken English. *English for Specific Purposes* 33, 66–76. doi: 10.1016/j.esp.2013.08.001.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning*, *67*(4), 959-997. https://doi.org/10.1111/lang.12253

Dang, T. N. Y. (2018a). A hard science spoken word list. *ITL-International Journal of Applied Linguistics*, *169*(1), 44-71.

Dang, T. N. Y. (2018b). The nature of vocabulary in academic speech of hard and soft-sciences. *English for Specific Purposes*, *51*, 69-83.

Dang, T. N. Y. (2019). The potential for learning specialized vocabulary of university lectures and seminars through watching discipline-related tv programs: insights from

medical corpora. *TESOL Quarterly*. https://doi.org/10.1002/tesq.552.\

Dang, T. (2020). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (ed.), *The Routledge Handbook of Vocabulary Studies (pp. 288-304*). New York: Routledge.

Dang, T. N. Y., Webb, S., & Coxhead, A. (2022). Evaluating lists of high-frequency words: teachers' and learners' perspectives. *Language Teaching Research*, *26*(4), 617-641.

Ha, A. Y. H., & Hyland, K. (2017). What is technicality? a technicality analysis model for eap vocabulary. *Journal of English for Academic Purposes*, *28*, 35-49.

Haruna, H. H., Ibrahim, B., Haruna, M., Ibrahim, B., & Yunus, K. (2018). Metadiscourse in students' academic writing: case study of Umaru Musa Yar'adua University and Al-Qalam University Katsina. *International Journal of English Linguistics*, *8*(7), 83-92. https://doi.org/10.5539/ijel.v8n7p83

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

Hu, M. & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–30.

Ibrahim, B., Haruna, H. H., Bashir, I., & Yunus, K. (2018). The usage of spatial prepositions in the headlines of major Nigerian newspapers. *International Journal of English Linguistics*, *8*(7), 13-22. https://doi.org/10.5539/ijel.v8n7p13

Janebi Enayat, M., & Derakhshan, A. (2021). Vocabulary size and depth as predictors of second language speaking ability. *System*, 99, 102521. https://doi.org/10.1016/J.SYSTEM.2021.102521.

Laufer, B. and Nation, P. (1995). Lexical richness in l2 written production: can it be measured? *Applied Linguistics, 16* (3), 307-322.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1), 15–30.

Laufer, B. (2013). Lexical thresholds for reading comprehension: what they are and how they can be used for teaching purposes. *TESOL Quarterly*, *47*(4), 867–872.

Lindqvist, C., Anna & Camilla, B. (2013). A new approach to measuring lexical sophistication in l2 oral production. In *L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis*, edited by Camilla Bardel, Christina Lindqvist, and Batia Laufer, 109–26.

Lu, C., & Coxhead, A. (2020). Vocabulary in traditional chinese medicine: insights from corpora. *ITL-International Journal of Applied Linguistics*, *171*(1), 34-61.

Macalister, J. (2016). Applying language learning principles to coursebooks. *English Language Teaching Today: Linking Theory and Practice*, 41-51.

Makino, Mark (2024). The purposes of first-year course syllabi according to corpus data, *EnglishUSA Journal*: Vol. 9, Article 5. Available at: https://surface.syr.edu/englishusa_journal/vol9/iss1/5

Miralpeix, I., & Muñoz, C.P. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching*, 56, (1), 1–24. https://doi.org/10.1515/iral-2017-0016.

Nation, I.S.P., (2001). Using small corpora to investigate learner needs: two vocabulary research tools. In: Ghadessy, M, Henry, A., Roseberry, R.L. (Eds.), *Small Corpus Studies and ELT*. John Benjamins Publishing, pp. 31–46.

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

Nation, I. S. P. (2011). Research into practice: *Vocabulary Language Teaching*, *44*(4), 529-539.

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

Nation, I. S. (2013). *Teaching & learning vocabulary*. Boston: Heinle Cengage Learning.

Nation, I. S. P. (2016). Making and using word lists for language learning and testing. *Making and Using Word Lists for Language Learning and Testing*, 1-224.

Ng, Y. J., Chong, S. T., Thiruchelvam, S., Chow, M. F., & Karthikeyan, J. (2020). Vocabulary threshold for the comprehension of malaysian secondary engineering texts as compared to the non-engineering genres. *International Journal of Innovation, Creativity and Change*, *14*(1), 488-504.

Nguyen C-D (2020). Lexical features of reading passages in english-language textbooks for vietnamese high-school students: do they foster both content and vocabulary gain? *RELC Journal*. DOI: 10.1177/0033688219895045

Nurmukhamedov, U., & Webb, S. (2019). Research timeline: lexical coverage and profiling. *Language Teaching*, *52*(2), 188–200.

Rabadi, R. I. (2023). Examining the role of breadth and depth of vocabulary knowledge in reading comprehension of english language learners. *Jordan Journal of Modern Languages and Literatures Vol*, *15*(1), 327-345.

Schmitt, N., Jiang, X., Grabe, W., (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, 95 (1), 26–43. doi: 10.1111/j.1540-4781.2011.01146.x.

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use english? Replication of van

Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, *50*(2), 212–226.

Sinclair, J. (1991). *Corpus, Concordance, Collocation: Describing English Language*. Oxford, UK: Oxford University Press.

Stamatović, M. V., Bratić, V., & Lakić, I. (2020). Vocabulary of L1 and L2 graduation theses written by English philology students: academic writing of Montenegrin and us students compared. *ELOPE: English Language Overseas Perspectives and Enquiries*, *17*(2), 101-113.

Sun, Y., & Dang, T. N. Y. (2020). Vocabulary in high-school EFL textbooks: Texts and Learner Knowledge. *System*, *93*, 102279.

Szudarski, P. (2018). *Corpus Linguistics for Vocabulary: A Guide for Research*. London

Tegge, F. (2017). The lexical coverage of popular songs in english language teaching. *System*, *67*, 87-98.

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

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in l1 and l2 listening comprehension: the same or different from reading comprehension? *Applied linguistics*, *34*(4), 457-479.

Vuković-Stamatović, M. (2020). Vocabulary complexity and reading and listening comprehension of various physics genres. *Corpus Linguistics and Linguistic Theory*, 16(3), 487-514. DOI: https://doi.org/10.1515/cllt-2019-0022.

Webb, S., & Rodgers, M. P. H. (2009a). The Lexical Coverage of Movies. *Applied Linguistics*, 30(3), 407–427.

Webb, S., & Nation, I.S.P. (2012). Computer-assisted vocabulary load analysis. in c.a. chapelle (Eds.), *The Encyclopedia of Applied Linguistics*, (pp. 1–10). Wiley Blackwell. https://doi.org/10.1002/9781405198431.wbeal0179.

Webb, S., & Paribakht, T.S. (2015). What is the relationship between the lexical profile of test items and performance on a

standardized english proficiency test. *English for Specific Purposes,* 38, 34–43.

Webb, S., & Nation, I. S. P. (2017). *How Vocabulary is Learned*. Oxford: Oxford University Press.

Webb, S. (2021). Research investigating lexical coverage and lexical profiling: what we know, what we don't know, and what needs to be examined. *Reading in a Foreign Language*, *33*(2), 278-293.

Yang, L., & Coxhead, A. (2022). A corpus-based study of vocabulary in the new concept English textbook series. *RELC Journal*, *53*(3), 597-611.

Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and l2 reading/listening comprehension: a meta-analysis. *Language Teaching Research*, 26(4), 696–725. https://doi.org/10.1177/1362168820913998.