Vocabulary Learning and Instruction
Volume 10, Issue 2, December 2021
http://vli-journal.org

A Comparison of Textbook Vocabulary Load Analysis

Stuart Benson[a] and Naheen Madarbakus-Ring[b]
[a]*University of Aizu;* [b]*Nagoya University of Commerce and Business*

Abstract

The popularity of using textbooks in second language programs in universities around the world continues to grow. Textbooks support teachers in their teaching by providing accessible materials and clear instruction. In addition, learners are guided by familiar lesson frameworks (e.g., beginning, middle, end) to guide their independent study (Swales, 1980). However, textbooks present many challenges. Learners' difficulties include the range in lexical knowledge they must possess (Nation, 2006) and the different lexical and grammatical features that are found in written textbook registers (Biber et al., 1998). This study investigates and outlines the vocabulary load of two English for Academic Purpose textbooks, using the British National Corpus and Corpus of Contemporary American English (BNC/COCA) 25,000 (Nation, 2012) and JACET8000 (JACET, 2016) word lists. The results show that for each textbook, more lexical demands are needed for second language learners in the JACET8000 compared with the BNC/COCA 25,000 lists. Understanding the content in textbooks will inform of the vocabulary-level requirements needed when taught in tertiary-level programs. Using a general and a Japanese-specific word list to identify possible pedagogical priorities can help to determine textbook priorities for teachers that can be applied to teaching in the Japanese classroom.

1 Background

For many tertiary-level institutions, the textbook is an integral component of any English for Academic Purposes (EAP) program. Richards (2001) observed how textbooks provide integral structure and syllabus guidance for teachers and learners in a language program. Specifically, textbooks offer instruction, guidance, and activities for learners to practise language through using communicative (e.g., role play, discussion activities) and linguistic (e.g., vocabulary definitions, gap fill) components (O'Loughlin, 2012).

Commercially, textbooks range from beginner to advanced levels that aim to become more difficult with each subsequent unit to accommodate learners in their language learning (Mares, 2003). Therefore, one important consideration for both teachers and learners is vocabulary. Coxhead et al. (2010) observed how "the reciprocal relationship between vocabulary knowledge and textbooks is critical" (p. 4). Sun and Dang (2020) concurred, suggesting that learners pay attention to both the textbook content and vocabulary to optimise learning.

Please cite this article as: Benson, S. and Madarbakus-Ring, N. (2021). A comparison of textbook vocabulary load analysis. *Vocabulary Learning and Instruction, 10*(2), 9–17. https://doi.org/10.7820/vli.v10.2.benson

Despite these observations, textbooks may lack the vocabulary knowledge necessary for learners and could present several problems for practitioners. As Macalister and Nation (2019) suggested, a language course needs to include high-frequency language items to help learners achieve the best possible vocabulary coverage. As a result, research focusing on textbooks has now turned to investigating the vocabulary loads in textbook content. Researchers have used two forms of vocabulary load analysis: the number of high frequency words included in textbooks and the number of word families needed to reach 95% and 98% coverage (Sun & Dang, 2020).

2 Previous Research

Two main studies have shown the first 2,000 word families cover 80% of a written text and 95% of spoken language. Eldridge and Neufeld's (2009) study found 1,400 of the first 2,000 words in the first four levels of the *Success* series. Similarly, O'Loughlin's (2012) study found 1,435 of the first 2,000 words in the first three levels of the *English File* series. These findings suggest that learners' vocabulary knowledge of the first 2,000 high-frequency words would account for 75% of the written discourse used in textbooks (Sun & Dang, 2020).

Table 1 shows studies that analyzed the number of word family knowledge needed to reach 95% and 98% coverage, respectively. Four of the five studies (Coxhead et al., 2017; Hajiyeva, 2015; Sun & Dang, 2020; Yang & Coxhead, 2020) found the vocabulary load of different commercial textbooks (CTs) reached 95% coverage by the first 3,000 words (including supplementary lists). Matsuoka and Hirsh (2010) found the *New Headway* series reached 95% coverage by the first 2,000 words while Yang and Coxhead (2020) found a higher level 4 of the *Yilin* textbook needed 4,000 words. Therefore, most textbooks achieve 95% coverage if learners have vocabulary knowledge of the first 3,000 word families. However, the same studies concluded different vocabulary knowledge needed by learners when considering the reach for 98% coverage. The studies reported varying vocabulary loads between the first 5,000 words (for the *New Concept English* series [Yang & Coxhead, 2020]) to the first 9,000 words (for the various *University Textbook* titles [Hajiyeva, 2015] and *Yilin* [Sun & Dang, 2020]).

Table 1. Studies Investigating the Vocabulary Load of Textbooks

Study	Textbook	Running words	95% coverage	98% coverage
Matsuoka and Hirsh (2010)	New Headway (Upper Intermediate)	44,877	95.5%–2,000	-
Harjiyeva (2015)	University Textbook (11 titles)	508,802	95%–3,500	98%–9,000
Coxhead et al. (2017)	English for Specific Purposes (ESP) Textbooks (15 titles)	380,078	95%–3,000	98%–7,000
Sun and Dang (2020)	Yilin (Four levels)	273,094	95.54%–3,000	98.02%–9,000
Yang and Coxhead (2020)	New Concept English (Two levels)	L3 -22,786 L4–18,109	95.59%–3,000 96.51%–4,000	98.08%–5,000 98.3%–6,000

These studies show that irrespective of the textbook title, the knowledge needed to reach 98% vocabulary coverage differs between textbooks. This suggests that further research is needed to analyze the vocabulary load in textbooks as the content may be too lexically demanding for the intended users.

3 Research Approach

We decided to investigate the vocabulary load of two EAP textbooks. The aim was to identify the vocabulary load of each textbook to determine its appropriateness to the student population it was being used for.

3.1 Materials

Two EAP textbooks were investigated for this study. The first textbook was an in-house textbook (IT), which was created by the teachers in an English department at a Japanese university. The book consists of eight units used over one academic year. The second was a commercial textbook (CT), which was developed by Cengage Learning, a reputable English language resources developer. The textbook consists of 12 units used over one academic year.

3.2 Context

Both books are considered appropriate to teach first year tertiary-level students based on their vocabulary and academic content. However, it is unclear about how appropriate these textbooks intended pre-intermediate levels correspond with the student populations' current proficiency scores. This led to formulating the following research questions:

(1) What is the vocabulary load of the in-house textbook (IT)?
(2) What is the vocabulary load of the commercially published textbook (CT)?
(3) How does the vocabulary load of each textbook compare when analyzed using a general word list and context-specific word list?

3.3 Method

The two chosen textbooks were scanned using Optical Character Recognition (OCR) Software. The scans were then cleaned by page and categorized by unit to prepare for the analysis. The data were run through the Range Program (Heatley et al., 2002) and the New Word Level Checker (Mizumoto, 2021). This study presents the preliminary results of the vocabulary load analysis of each textbook. The analysis used two word lists: a general word list and a context-specific word list. Nation's (2012) BNC/COCA base words lists were utilized as the general word list. The BNC/COCA divides the most frequent 25,000 word families into 25, 1,000 word base lists, according to their frequency. In addition, Nation's (2012) supplementary word lists (i.e., proper nouns, marginal words, transparent compounds, abbreviations) were also used. The New JACET8000 word list (JACET, 2016) was utilized as the context-specific word list. The New JACET8000 word list

is an 8,000 lemma educational word list, created for Japanese learners of English, specifically tertiary-level learners. The New Word Level Checker (Mizumoto, 2021) divides the New JACET8000 word list (JACET, 2016) into eight, 1,000 word base lists, according to their frequency. Nation's (2012) BNC/COCA base word list was used due to its size, and with the Range program, the ability to analyze vocabulary at various frequency levels. The New JACET8000 word list (JACET, 2016) was used as it contains the most frequent vocabulary within the Japanese tertiary level, which is the context where each textbook is utilized.

4 Results

Table 2 presents the vocabulary load of the IT analyzed using Nation's (2012) BNC/COCA word lists. Nation's (2012) supplementary lists were included in the vocabulary load analysis when considering the 95% and 98% thresholds because Nation (2013) pointed out that once known, these words are not a burden to learners. In addition, due to their high coverage, the words in the supplementary lists were important to achieve 95% or 98% coverage.

As illustrated in Table 2, with the supplementary lists, the IT reached 95% coverage between the 2,000 and 4,000 base word families and reached 98% coverage between 4,000 and 7,000 base word families. Therefore, if learners know the first 7,000 word families plus the four supplementary lists, they could theoretically comprehend the IT. However, upon further analysis of Table 3, Unit 1 and 2 had varying levels of coverage that would affect comprehension of those units.

Table 3 shows the CT reached 95% coverage between the 2,000 and 3,000 base word families and reached 98% between 3,000 and 6,000 base word families. As with Table 2, Table 3 also shows some indiscretions in coverage within the textbook. For example, the vocabulary coverage in Unit 4 will be difficult for learners, compared with that of the other 11 units in the textbook. Compared with the results of the IT however, the ranges for each unit are similar, with few outliers.

Table 2. Cumulative Coverage of the IT by Nation's (2012) 25,000 BNC/COCA Word Lists and Supplementary Lists

Nation (2012) BNC/COCA + supplementary lists	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6	Unit 7	Unit 8	Total
Supplementary lists (31–34)	2.28	1.46	1.8	1.45	8.31	2.83	1.96	0.99	2.36
1	87.55	81.61	83.41	87.96	84.38	79.12	81.85	83.54	83.78
2	**95.16**	91.86	90.15	93.83	93.74	92.65	90.96	93.05	92.72
3	**98.44**	94.76	**95.16**	**98.39**	**96.36**	**97.40**	**96.85**	**96.76**	**96.71**
4	98.86	**95.51**	96.54	98.71	97.13	**98.40**	97.91	**98.13**	97.54
5	99.39	96.23	**98.12**	99.95	**98.39**	99.06	**98.62**	98.98	**98.39**
6	99.69	97.41	98.68	99.70	98.67	99.42	99.26	99.39	98.93
7	99.76	**98.66**	98.94	99.88	98.98	99.52	99.42	99.45	99.31
8	99.81	99.22	99.14	99.90	99.22	99.55	99.56	99.51	99.50

Note: Bolded items are coverage thresholds.

Table 3. Cumulative coverage of the Commercial Textbook by Nation's (2012) 25,000 BNC/COCA Word Lists and Supplementary Lists

Nation (2012) BNC/COCA + supplementary lists	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6	Unit 7	Unit 8	Unit 9	Unit 10	Unit 11	Unit 12	TOTAL
Supplementary lists (31–34)	3.87	5.51	6.87	4.27	4.91	4.71	4.17	4.44	6.85	4.61	2.94	3.97	4.62
1	84.89	88.5	85.9	83.74	85.54	87.48	88.28	84.23	84.48	86.28	84.95	84.74	86.00
2	**95.19**	**96.89**	93.43	93.76	94.34	94.81	**95.33**	92.00	92.98	94.48	**95.06**	92.69	94.47
3	97.24	**98.64**	**97.61**	**96.27**	**97.85**	**97.84**	97.41	**96.1**	**97.17**	**97.19**	**98.22**	**97.35**	**97.49**
4	97.71	99.12	**98.41**	97.24	**98.25**	**98.65**	**98.31**	97.35	97.96	**98.00**	98.62	**98.53**	**98.23**
5	**98.62**	99.63	99.27	97.97	99.04	99.11	98.89	**98.33**	**98.58**	99.24	99.15	98.87	98.91
6	98.72	99.87	99.44	**98.67**	99.59	99.57	99.27	98.76	99.20	99.63	99.37	99.15	99.29
7	98.99	99.90	99.71	98.97	99.83	99.77	99.37	98.91	99.34	99.66	99.56	99.34	99.45
8	99.09	99.93	99.71	99.15	99.92	99.83	99.43	99.19	99.48	99.75	99.81	99.43	99.57

Note: Bolded items are coverage thresholds.

Table 4. Cumulative Coverage of the IT by the New JACET8,000 (JACET, 2016)

Base word lists	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6	Unit 7	Unit 8	TOTAL
Proper nouns	2.59	3.02	2.33	1.89	4.05	3.90	6.22	2.33	3.46
1	85.55	79.96	83.97	82.98	78.87	78.24	82.56	83.97	81.88
2	94.45	88.91	93.51	92.92	90.27	90.56	92.60	93.51	91.68
3	**97.12**	93.22	**96.96**	**95.38**	93.20	**96.00**	**95.19**	**96.96**	**95.36**
4	**98.26**	94.55	**98.08**	96.77	94.85	97.25	96.67	**98.08**	96.67
5	98.60	94.97	98.39	97.97	**95.64**	97.88	97.66	98.39	97.24
6	99.05	**96.21**	98.54	**98.13**	96.36	**98.13**	**98.03**	98.54	97.76
7	99.15	96.46	98.59	98.25	96.67	98.45	98.27	98.59	97.97
8	99.28	97.35	98.74	98.29	96.98	98.78	98.55	98.74	**98.30**

Note: Bolded items are coverage thresholds.

Table 4 shows the vocabulary load analysis results of the IT using the New JACET8000 (JACET, 2016). The IT reached 95% coverage between 3,000 and 6,000 and reached 98% coverage between 4,000 and 8,000. Table 4 highlights Unit 2 and Unit 5 not reaching 98% coverage. Comparing these units alongside Units 4, 6, 7, and 8 indicates that a wide range of coverage will affect comprehension for Japanese tertiary-level learners.

Table 5 shows that the CT reached 95% coverage between the 2,000 and 5,000 base word lists and reached 98% coverage between 5,000 and 8,000 base word families. As with the IT, CT Units 5, 6, 8, 11, and 12 did not reach 98% coverage.

5 Discussion

When comparing the results of Tables 2 and 3 with Tables 4 and 5, it is clear that both textbooks are lexically demanding for Japanese tertiary-level learners. If the books were solely analyzed using a general word list, such as Nation's (2012) BNC/COCA, the results would not illustrate this problem. Therefore, the preliminary results further highlight the need for using context-specific word lists.

Ideally, the lexical coverage in each unit of a textbook should be identical or show incremental increases. This is not the case for both; with the results highlighting indiscretions within each textbook that could affect the comprehension of certain units. This, however, is somewhat unsurprising in the IT, as it was created by several people and presumably, the vocabulary profile was not analyzed. This issue can be easily fixed as the materials can be redesigned following the results of this study. However, a university using the CT will need supplementary material, such as word lists, to assist learners when studying certain units.

6 Implications for Future Research

These preliminary results point toward further analyses in several areas. Firstly, the analysis could focus on other word types (e.g., academic vocabulary) to understand the different types of words that are included in the vocabulary load of each textbook. Secondly, an analysis of the incremental vocabulary increase per unit could indicate the changes in the level of difficulty between the

Table 5. Cumulative Coverage of the CT by the New JACET8000 (JACET, 2016)

Base word lists	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6	Unit 7	Unit 8	Unit 9	Unit 10	Unit 11	Unit 12	TOTAL
Proper nouns	6.54	8.53	7.81	7.17	8.22	7.88	8.03	8.05	11.23	7.83	7.59	9.20	8.24
1	85.86	88.26	87.29	81.76	84.05	85.75	86.65	83.75	83.69	86.70	83.21	85.92	85.26
2	93.92	**95.58**	94.56	91.01	92.81	92.62	93.48	91.80	92.48	94.04	92.71	92.77	93.17
3	**96.35**	97.16	**96.50**	94.26	**96.09**	**95.57**	**95.83**	94.51	94.82	**96.89**	**95.49**	**96.04**	**95.82**
4	97.67	97.93	97.29	94.96	97.46	96.55	97.03	**97.11**	**95.93**	97.93	96.77	97.26	97.01
5	97.90	**98.05**	97.81	**95.70**	97.69	96.91	97.28	97.40	96.72	**98.43**	97.25	97.38	97.40
6	**98.09**	98.16	**98.15**	96.11	97.99	97.13	97.56	97.73	97.07	98.49	97.46	97.56	97.65
7	98.19	98.28	98.24	96.49	98.25	97.38	**98.21**	97.82	97.72	98.63	97.73	97.77	97.92
8	98.36	98.42	98.42	97.28	98.55	97.86	98.24	97.96	97.96	98.75	97.79	97.83	**98.15**

Note: Bolded items are coverage thresholds.

textbooks. These results could show if each unit increase is appropriate for learners using the textbook. Finally, using these results, supplementary materials could be designed to scaffold vocabulary learning for learners. In turn, these materials can help bridge any potential learning gaps between the textbook vocabulary and learners' vocabulary knowledge.

7 Conclusion

This article presented the main vocabulary load findings from an IT and CT using a general and Japanese-specific word list. The initial results show that learners have higher lexical demands when considering vocabulary knowledge in JACET8000 compared with the BNC/COCA 25,000 general word list. These preliminary results will be analyzed further to determine the type of words, word families and lemmas used by skill and by unit in each textbook. Although textbooks are useful guidance in helping learners in second language classrooms, there is some evidence that there may be a discrepancy between the textbook context and learners' vocabulary knowledge. Therefore, further analysis is needed to identify the textbook's vocabulary load so that teachers can assist learners in reaching the vocabulary coverage.

References

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

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. https://doi.org/10.1016/j.jeap.2017.11.001

Coxhead, A., Stevens, L., & Tinkle, J. (2010). Why might secondary science textbooks be difficult to read? *New Zealand Studies in Applied Linguistics, 16*(2), 37–52. https://doi.org/https://search.informit.org/doi/epdf/10.3316/informit.893694181605406

Eldridge, J., & Neufeld, S. (2009). The graded reader is dead, long live the electronic reader. *Reading, 9*(2), 224–244. https://doi.org/10.1.1.586.6442

Hajiyeva, K. (2015). A corpus-based lexical analysis of subject-specific university textbooks for English majors. *Ampersand, 2*, 136–144. https://doi.org/10.1016/j.amper.2015.10.001

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *Range computer programme*. https://www.victoria.ac.nz/lals/about/staff/paul-nation

JACET. (2016). *Jacet8000*. http://www.j-varg.sakura.ne.jp/publications/

Macalister, J., & Nation, I. P. (2019). *Language curriculum design*. Routledge.

Mares, C. (2003). Writing a coursebook. In B. Tomlinson (Ed.), *Developing materials for language teaching* (pp. 130–140). Continuum.

Matsuoka, W., & Hirsh, D. (2010). Vocabulary learning through reading: Does an ELT course book provide good opportunities? *Reading in a Foreign Language, 22*(1), 56–70. https://doi.org/10.20581/arele.31.0_49

Mizumoto, A. (2021). *New word level checker* [Web application]. https://nwlc.pythonanywhere.com/

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. (2012). The BNC/COCA word family lists (17 September 2012). *Unpublished paper.* www. victoria. ac. nz/lals/about/staff/paul-nation

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

O'Loughlin, R. (2012). Tuning in to vocabulary frequency in coursebooks. *RELC Journal, 43*(2), 255–269. https://doi.org/10.1177/0033688212450640

Richards, J. C. (2001). *The role of textbooks in a language program*. Cambridge University Press. https://doi.org/https://www.professorjackrichards.com/wp-content/uploads/role-of-textbooks.pdf

Sun, Y., & Dang, T. N. Y. (2020). Vocabulary in high-school EFL textbooks: Texts and learner knowledge. *System, 93*, 102279. https://doi.org/10.1016/j.system.2020.102279

Swales, J. (1980). ESP: The textbook problem. *The ESP Journal, 1*(1), 11–23. https://doi.org/10.1016/0272-2380(80)90006-2

Yang, L., & Coxhead, A. (2020). A corpus-based study of vocabulary in the new concept English textbook series. *RELC Journal*, 2020, 1–15. https://doi.org/10.1177/0033688220964162