What's in Your Textbook? An Analysis of the Vocabulary in a Second Language Learning Textbook

Mark Alberding, Asia University

ABSTRACT

Once only the province of computational linguists, corpus research is now within easy reach of any non-specialist with basic computer skills and a research question. One research area that is particularly accessible is that of vocabulary teaching and learning. This paper explains how teachers can analyze the vocabulary in their textbook with an uncomplicated computer program, then use the results to inform their decisions for vocabulary teaching when using the textbook.

INTRODUCTION

In many English courses in Japanese universities, teachers are required to use a commercially produced textbook, which can be expected to be the foundation for instruction in the course and thus the principal source of the vocabulary items and language structures and functions students will encounter. Of course, teachers often employ additional materials to supplement their textbooks to provide students with further practice of target vocabulary, structures and functions. In the case of the latter, it is a simple matter of consulting the textbook to see which are featured in a given lesson and then planning supplemental materials accordingly. On the face of it, it would seem similarly straightforward to determine what vocabulary is in a particular textbook unit: the teacher can identify important words during the course of preparing the unit or consult the textbook's vocabulary list for the unit, a feature of most current textbooks.

But such measures alone are insufficient for teachers to make the best decisions regarding vocabulary instruction in their classes. In order to make well-informed decisions about vocabulary teaching for their individual situations, teachers need to have the answers to some fundamental questions about the vocabulary content of their textbooks, such as "is the level of the vocabulary in this textbook appropriate for my learners?", "what vocabulary should I expend class time on?", and "which vocabulary items should be given additional exposure and practice through supplemental materials"? A good starting point for answers to these questions is a basic analysis of the vocabulary in the textbook being used, as will be demonstrated in this paper.

Level of Difficulty of the Vocabulary in a Textbook

Because the textbook is almost certainly going to be the primary source for vocabulary in a course, teachers should have an objective assessment of the difficulty of its vocabulary. The learner-level of the textbook as advised by the publisher is a good starting point in making such an assessment, but it is by necessity only an overall indicator. When determining the level of a textbook, publishers consider many features in addition to the difficulty of the vocabulary, such as the syntactic complexity of the language used, the difficulty of the grammatical constructs presented, and the length of both reading and listening texts.

Many textbooks include some sort of listing of the principal vocabulary items encountered in the textbook and this can help determine the suitability of the vocabulary for a given set of learners. However, the contents of such lists still require analysis to be useful—intuitions of difficulty arising from simply perusing the lists is insufficient. An analysis of all the words present in the textbook, not just those in the vocabulary lists, will provide a complete picture of the vocabulary load of the textbook. Teachers can see how many and which words are in the first one or two thousand most frequent words of English, for example, or which low-frequency words are present and if there are too many of these for the level of the learners. Such data will help inform the teacher whether or not the vocabulary load in the textbook is appropriate for the learners at that level.

Distribution and Repetition of Vocabulary

Research shows that it is necessary to encounter a word in a variety of contexts a number of times, at regular intervals, in order for the learner to have a realistic chance of learning the word (Nation, 2001, Schmitt, 2000). An analysis of the textbook can show if target vocabulary occurs frequently enough and is given enough repetitions over time to provide optimum vocabulary-learning conditions. The results can guide teachers in deciding how best to supplement the text with activities that will give learners exposure to target vocabulary that is not sufficiently presented in the textbook.

ANALYZING THE TEXTBOOK

Creating The Textbook Corpus

The first step in an analysis of the vocabulary in a textbook is to create a corpus consisting of the contents of the textbook. For this investigation, selected contents of the textbook *Firsthand Success: Beginners' Course 2, Gold Edition* (Helgesen, Brown, & Kahny, 2001), were converted into machine-readable ASCII format text files, so that they could be "read" by the computer program used to analyze them. This was done by entering the contents into files using word processing software (Microsoft Word 2000, Microsoft Corporation, 1999). Scanning of text using Optical Character Recognition (OCR) software is also possible, but the great variety of formatting found on a page of a typical modern textbook creates significant accuracy problems for off-the-shelf OCR software, such that the time spent correcting errors and omissions in the scanned text is greater than the time it takes to type in the text itself.

The corpus created for this investigation consisted of the textbook's 11 principal units, the 2 review units, the 10 writing units, and the 10 learning summaries, which contained the key structures, functions, and vocabulary for each unit. In the judgment of the author, the material included in the corpus comprised what would typically be thought of as the core instructional material for the textbook, that which most teachers could reasonably be expected to cover when using the textbook for a year-long university English course. All the text from the selected material was entered into the corpus including unit titles, section headings, and instructions, since students would encounter it all during their use of the textbook.

Omitted from the corpus were the credits, acknowledgements, introduction (for teachers only), table of contents, *Resources* page explaining support materials such as the CD and website, two pages of *Extra Vocabulary* consisting of days, months, numbers, and countries and nationalities, and a page called *Tool Box* on the last page of the textbook which contained some of the classroom language found in several places elsewhere in the textbook. The omitted pages were considered to contain ancillary material that not all teachers would be likely to use.

The RANGE Computer Program

The textbook corpus was analyzed using the vocabulary analysis program *RANGE* (Heatley, Nation, & Coxhead, n.d.), freeware available at http://www.vuw.ac.nz/lals. *RANGE* is a powerful yet easy-to-use program that allows the user to analyze a number of features of vocabulary in multiple texts simultaneously. As described in the instructions accompanying the program, *RANGE* "provides a range or distribution figure (how many texts the word occurs in), a headword frequency figure (the total number of times the actual headword type appears in all the texts), a family[1] frequency figure (the total number of times the word and its family members occur in all the texts), and a frequency figure for each of the texts the word occurs in" (p.1). *RANGE* can also be used to compare the contents of a corpus with word lists, to see which words are present in the lists. Although the user can make custom word lists, the program comes with three word lists, which were used for the analysis described in this paper. Lists one and two contain, respectively, the first and second 1000 most frequent words of English from *A General Service List of English Words* by Michael West (Longman, London, 1953); list three is the *Academic Word List* (Coxhead, 2000) and contains words that are not in the first 2000 words but are frequently found in academic texts. All three lists contain the base forms of words (the *headword type* mentioned above) as well as their family members.

RANGE can examine up to thirty-two different text files simultaneously. If all the textbook units described above had been placed in individual files the total number of files would have been thirty-three. A central point of this investigation was to make judgments about the vocabulary in the textbook based on the distribution of words across all the units of the textbook, and since each unit could not be placed in its own file without exceeding the capacity of the program, no ideal solution could be arrived at. It was decided to combine the two review units into a single file, for a total of thirty-two files. This allowed for keeping all of the main and writing units, and the learning summaries in individual files, and provided the maximum number of texts for coverage results within the parameters of the program.

RESULTS

The summary results of running *RANGE* over the textbook corpus is shown below (Table 1). The corpus consisted of 13058 tokens, or running words[2] comprised of 1407 types.[3] The summary data show coverage of the corpus by the words in the three word lists and those words not in the lists. For example, the first line shows that 10731 running words were found in the list of the first 1000 words of English, and that these made up 82.2% of all running words in the corpus, the 798 types found in the list made up 56.7% of all types in the corpus, and 506 word families were represented.

[1] A word family consists of a headword, all its inflected forms, and some of its derived forms. For example, for the headword *wonder*, an inflected form is *wonders* (v) and a derived form is *wonderful*.

[2] A token, or running word, is a word in a text. A count of tokens is a count of the total number of words in a text, including repetitions of the same word.

[3] A type is each different word in a text. When counting types, each type is counted only once per text.

Table 1
Summary Data of Textbook Corpus Analyzed with the *RANGE* Program

Word List	Tokens / %	Types / %	Families
1st 1000	10731/82.2	798/56.7	506
2nd 1000	1093/ 8.4	214/15.2	162
Academic	188/ 1.4	34/ 2.4	30
Not in the lists	1046/ 8.0	361/25.7	N/A
Total	13058	1407	698

DISCUSSION

Coverage

The summary data show that a little more than 82% of the corpus was covered by the first 1000 words of English and about 91% was covered by the first 2000 words. This means that 82% and 91% of the running words of the textbook are found in the first 1000 and 2000 words of English, respectively. When compared with coverage figures for different genres of English (Table 2), these numbers compared favorably with the language found in conversation and fiction and less favorably with that found in newspapers and academic texts. In fact, the coverage of the textbook by the first 2000 words was almost identical to the coverage of conversational language, though the distribution was slightly different. This was an appropriate result, given that the focus of the textbook is on developing speaking and listening skills. Students can comfortably negotiate much of the textbook within the confines of the first 1000 words of English, which would seem to make it a suitable text for low-level students who have had prior study experience studying English, such as those in the author's class.

Table 2
Coverage of Different Genres of English by Word Level

Word Levels	Textbook Corpus	Conversation	Fiction	Newspapers	Academic Text
1st 1000	82.2%	84.3%	82.3%	75.6%	73.5%
2nd 1000	8.4%	6%	5.1%	4.7%	4.6%
Academic	1.4%	1.9%	1.7%	3.9%	8.5%
Other	8.0%	7.8%	10.9%	15.7%	13.3%

Note. Genres data from Nation (2001, p. 17).

Suitability of the Vocabulary Level of a Textbook for a Specific Group of Learners

The simplest way for teachers to determine if a textbook's vocabulary level is suitable for a specific group of learners is to look at the distribution of tokens and types amongst the lists and words not in the lists. We have see that over 90% of the running words in the textbook fell within the first 2000 words of English. The 2000 words level is frequently mentioned as the basic initial goal of many second language learners (Schmitt, 2000, p. 142).

To determine whether this is a realistic goal for a group of learners it would be very useful to have an objective measure of the learners' existing vocabulary knowledge. In the absence of results from a validated vocabulary test for the specific group of learners using the textbook examined here, conjecture about their vocabulary knowledge remains only that. However, the data from such a test (Beglar & Hunt, 1999) the author conducted for 52 learners in the same program, but who had placed at a level significantly higher than those using the textbook, show that approximately 31% of the learners tested at between 500 and 800 words, and approximately 72% tested at 1200 words or fewer. It is reasonable to assume that the majority of the lower-level learners would not have a higher word knowledge level and would probably have one substantially lower than the higher-level students. Further evidence of the vocabulary level of the learners using the textbook comes from records of the learners' year-long extensive reading program which shows that 85% of the students were reading comfortably at the 400 words level and the remaining 15% were reading only at the 600 words level.

This would suggest that studying in the environment provided by the textbook would provide a significant but not overwhelming challenge for students with between 400 and 600 words of existing knowledge. Since 82.2% of the textbook was covered by the first 1000 words and only 8.4% was covered by the next 1000, the teacher could choose to focus vocabulary learning time for each class on those words of a unit that appear in the first 1000 words. This information is available in the detailed results provided by *RANGE* (see Appendix A for an excerpt of the complete results). Once the results have been entered into a spreadsheet program it is a simple matter to sort the data to show, for example, which words are present in a given unit of the textbook.

Vocabulary Load: Individual Words

The numbers for coverage of the text by running words provides a broad measure of the vocabulary load of the textbook. Teachers are also interested in knowing how much different vocabulary is presented in a textbook, and so would need to know the number of types in the textbook. This gives a better idea of the vocabulary load of the textbook and more clearly points the direction for supplementing the textbook vocabulary.

The summary data (Table 1) shows a total of 1407 types, 798 of which appear in the first 1000 words. Yet this is not the nearly 80% of the first 1000 words that it might seem. As mentioned above, *RANGE* uses word families for its lists, and the list of the first 1000 words used by the program includes around 4100 *types* for 1000 *word families,* not simply 1000 *types*. Thus to get a more accurate idea of how many of the first 1000 words of English are covered by the textbook, it is necessary also to look at how many word families from the first 1000 are represented in the text. This number is 509; in other words, the textbook provides exposure to members of half of the word families in the first 1000 words of English.

For a 40-50 hour course, as this textbook is intended, this would seem to be a fairly small number. If the learning goals of the course include increasing the students vocabulary to the point at which they know even just the first 1000 words of English, the text would appear to be inadequate as a primary source for vocabulary instruction. In order to make decisions about what supplemental vocabulary to include in the course, the teacher would likely want to know exactly what vocabulary from the first 1000 words was *not* present in the textbook. To find this, he or she could compare the words from the textbook identified by *RANGE* as present in the first

1000 words with the first 1000 word list used by *RANGE* and then make decisions about supplementary vocabulary accordingly.

Vocabulary Distribution and Learning Goals

One of the most valuable functions of *RANGE* is to provide figures for distribution or range of all words in the corpus across all the texts it is run on. It provides summary data for distribution of words in the texts (Table 3), and detailed results for all the words in the corpus (Appendix A).

Table 3
Summary of Distribution/Range of Types in the Corpus

This Number of Words	Appears in this Many Texts	This Number of Words	Appears in this Many Texts	This Number of Words	Appears in this Many Texts
523	1	13	11	3	21
326	2	27	12	8	22
174	3	9	13	2	23
97	4	9	14	1	24
46	5	3	15	1	25
42	6	8	16	2	27
24	7	10	17	1	28
16	8	1	18	3	29
15	9	3	19		
33	10	5	20		

Note. No words appeared across 32, 31, 30, or 26 texts, so these numbers of texts do not appear in the data.

The summary data show that very few words appear in more than one or two units in the textbook. 38% of the types (523) appear in only one unit and 73% of types (1023) appear in three or fewer units. In and of itself, these numbers are not unusual: Nation (2001, p.67) cites Kucera and Brown's (1982) finding that a learner at the 1000 word level would, on average, have to read or listen to 10,000 running words of academic texts in between meetings of the same word. Nevertheless, their finding was for standard texts, not Second Language Learning textbooks, which could reasonably be expected to provide, by design, more frequent encounters with words deemed important for learners to acquire. Barely 10% of the types (142) in this textbook occur in ten or more units, which is a very small number of repeated encounters for vocabulary learning purposes.

While the summary data provide an overall view of vocabulary range, it is also necessary to look at the detailed results of type distribution for the complete picture regarding encounters with and repetitions of a word. In this case, one needs to know the number of repetitions as well as the range of repetitions across the units of the textbook, in order to determine if the word is being suitably recycled in the textbook. Repetition is an important part of vocabulary learning and it is unlikely that students will remember a word encountered only a few times during a year-long course. Nation (2001, p.76) cites extensive memory research, including that specific to second language learning, showing that spaced repetition of vocabulary results in better retention rates than a large amount of repetition in a limited time period, such as multiple repetitions in one class period.

Accordingly, teachers would want to know, for example, whether a word that appears in fifteen textbook units with an overall frequency of thirty-five occurs in fairly regular, spaced repetitions, as shown in Fig. 1 or if it occurs in massed repetitions a limited number of times, as shown in Fig. 2. The kind of repetition shown in Fig. 1 is considered to provide much better circumstances for vocabulary learning than the kind of repetition shown in Fig. 2

Unit No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Frequency	6	4	0	6	0	3	5	0	2	4	0	2	0	3	0

Figure 1. A representation of regular, spaced repetitions of a word in the course of multiple units.

Unit No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Frequency	12	0	0	0	0	0	0	0	0	10	0	0	0	0	13

Figure 2. A representation of massed repetitions of a word in the course of multiple units.

A full discussion of the detailed results for all words is beyond the scope of this paper, so a narrow selection of the data—content words with a range of 10 or more—will be examined to illustrate the value of the detailed results in determining range and repetitions of words. 116 types occur in ten or more texts in the corpus. Of these, fifty-six, or 48% are content words, which is within the standard ratio of English for function to content words. The thirty-eight content words with the greatest range and frequency are shown in Appendix B, in a table which has been coded to show repetitions. The figures are very good for these words in respect to their range in conjunction with frequency of repetition. Almost all words appear in every main unit at least two or three times, which would satisfy general conditions of spaced repetition. However, closer examination of the words themselves reveals that almost all have to do with instructions or other classroom language, which is confirmed by an examination of these words in the textbook. There are few instances of these words being used for anything other than instructions or headings, with the exception of *like*.

This is not a negative result: these words are very important for students to function successfully in the classroom, and many of them are used frequently in non-pedagogical environments. Clearly, it is important for students to learn these words and there is little doubt that with their range and repetitions, these words will be learned well before the end of the course. Nevertheless, one would expect such words to be very well represented throughout a textbook; what of the words that have less to do with classroom language?

All thirty of the content words with a range of six are shown in Appendix C, in a table coded to show repetitions identically to the table in Appendix B. Few words in this table have primarily to do with classroom language. Rather, they are usually considered as general service vocabulary, and represent a different kind of vocabulary learning goal. As seen earlier, there are relatively few types overall in the textbook, so it is particularly desirable that the general service content words receive a good number of spaced repetitions.

The table in Appendix C shows that while the range and repetition of types in the table have less range and fewer repetitions than those in the table in Appendix B, some have spaced repetitions of three or more with a frequency of one or two in each instance, for example *years*, *friend*, *great*. However, it is clear that most words in Appendix C have limited spaced repetitions despite their range of at least six. In fact, this is the case for almost all general service words in the textbook: they have very limited repetitions, which decreases the likelihood of their retention. Thus it is essential that the teacher identify the general service content words in the lower ranges, perhaps eight and below, and include them in supplemental materials if they are to have a good chance of being learned during the course of the year.

Low-Frequency Words

Only 1.4% of running words and 2.4% of types in the corpus fell within the academic word list used by *RANGE*. This is appropriate, as the textbook was being used by learners not expecting to function in an English speaking academic environment. In order to decide if any of these are worth spending time with it is again necessary to look at the range and frequency of the items.

Table 4
The 10 Most Frequent Low-Frequency Words in the Corpus

Type	Range	Frequency	Type	Range	Frequency
Partner	13	39	Summary	10	10
Challenge	11	13	Job	6	16
Clarification	10	10	Computer	5	9
Functions	10	10	Team	4	7
Similar	10	11	Computers	3	3

None of the ten most frequent words (Table 4) had a range greater than 13 and all six of the words with a range of 10 or greater were again part of instructions (e.g. *partner* and *similar*), or headings (e.g. *functions* and *clarification* occur only as headings in the learning summaries). Since these words are going to be encountered frequently by students in the course of using the textbook, the teacher will want to make sure the students know the meaning of these words but will be aided in this by their range and repetition.

An examination of the academic words in the complete results for the corpus shows that 40% of the low-frequency words are classroom or pedagogical words and only a few of the remaining 60% occur in more than one unit in the textbook (*jobs, computer, computers, team, construction, designer, intelligent)*. These can be dealt with as they arise, most profitably with a direct first language translation if possible. Such treatment notwithstanding, if students have not already encountered these words, it is unlikely they will learn them, as they will encounter them only once or twice in the course of the entire year in the absence of supplementary materials which would recycle them several times. The effort the teacher would have to expend to ensure that these low-frequency words were recycled at regular intervals in supplementary materials is probably not worth the effort, but the teacher would nonetheless want to check the list of words for any exceptions.

Words Not in the Lists

The 8% of running words not in any of the lists might raise some alarms, as the number of very low-frequency words is almost the same percentage as the 8.4% of words found in the second 1000 words. Yet, an examination of the most frequent of these words reveals a mixed bag of familiar words that are nonetheless infrequent in everyday usage (e.g., *movie, hometown, jacket*) classroom/pedagogical words (e.g., *adjectives*, *phrase*, *pronunciation*, the affix *-ing*), and words specific to the textbook context (e.g., the section headings *Duet* and *Solo*). Given the existing challenges of adequately supplementing the first 1000 words, it is likely that the teacher will want to treat these similarly to the academic words as described in the previous section of this paper.

CONCLUSION

Certainly, one cannot expect a four-skills textbook to provide extensive vocabulary practice, yet one would hope that the design of the textbook would be such that the vocabulary items it does contain are recycled at regular intervals in order both to refresh the students' learning and give them needed practice of the words. Because the 2000 word level is an essential goal for second language learners, it would be hoped that the textbook would enable the students to attain at least the 1000 word level by the end of the course.

The analysis of the vocabulary in the textbook used here showed that while the overall contents of the textbook reflected a level of difficulty appropriate for the group of learners using it, and its coverage by the first 2000 words of English compared favorably with spoken language, at 798 types within 500 word families in the first 1000 words of English, the textbook does not contain adequate vocabulary items for a year-long course, nor does it provide enough spaced repetition of the items it does contain. Although it is comforting to think that because the textbook provides a list of important vocabulary for each unit the students will learn this vocabulary in the course of using the textbook, the reality does not match this perception. An analysis of range and repetition, just a limited portion of which has been demonstrated and discussed here, shows that it is necessary to extend the practice of most of the vocabulary in the textbook if the students are to have a good opportunity to learn and retain it during the course. Furthermore, without significant supplementation by the teacher to extend the vocabulary of the course beyond that offered in the textbook, students probably have little chance of increasing their vocabulary to even the 1000 word level.

The analysis conducted in the paper was not intended to criticize the vocabulary of a particular textbook; indeed, the author assumes most textbooks suffer from similar issues as those described above. Is it possible to write a textbook that not only mostly uses words within the first 1000 or 2000 words but uses most if not all of the first 1000 words and gives them well-spaced repetitions? The answer is almost certainly "no,"; teachers must work within the imperfect confines of existing textbooks. Still, as this paper demonstrates, teachers can reduce some textbook limitations by becoming familiar with the vocabulary contents of their textbooks. Teachers generally have a limited amount of time that they can devote to vocabulary-focused instruction, i.e. formally presenting vocabulary items and concomitantly providing activities that are designed to practice and reinforce the target vocabulary. Furthermore, because the number of vocabulary items in any language is so large, only a small number can be dealt with through vocabulary-focused instruction and most will either have to be acquired through exposure or not at all (Schmitt, 2000, p.3). These factors make it very important for teachers to be well-informed about the vocabulary in their textbooks, so that that they can make pedagogically-sound decisions regarding vocabulary instruction. Although it is time consuming to perform a vocabulary analysis of one's textbook, the results show the teacher which and how many words need to receive additional practice beyond the textbook to meet the vocabulary learning goals of the course.

References

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing* 16 (2), 131–162.

Coxhead, A. (2000). A new *Academic Word List*. *TESOL Quarterly* 34 (2), 213-218.

Heatley, A., Nation, P., and Coxhead, A. (n.d.). *RANGE* (Version 29b) [Computer Software]. Wellington, N.Z.: School of Linguistics and Applied Language Studies, Victoria University.

Helgesen, M., Brown, S., & Kahny, J. (2001). *Firsthand success: Beginners' course 2, gold edition*. Hong Kong: Longman Asia ELT.

Microsoft Corporation (1999). *Microsoft Word 2000*. [Computer Software]. Redmond, WA.: Microsoft Corp.

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

Nation, P. (2002). *RANGE and FREQUENCY programs for Windows based PCs*. [Computer Software Instructions]. Wellington, N.Z.: School of Linguistics and Applied Language Studies, Victoria University.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.