

What's in your textbook? An analysis of the vocabulary in a second language learning textbook

Mark Alberding

Asia University, Tokyo

Reference data:

Alberding, M. (2006). What's in your textbook? An analysis of the vocabulary in a second language learning textbook. In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (Eds.) *JALT2005 Conference Proceedings*. Tokyo: JALT.

Corpus research, which was once only a province of computational linguists, is now within easy reach of any non-specialist with basic computer skills. A research area that is both accessible and has concrete pedagogical applications is that of vocabulary teaching and learning. This paper discusses the results of the author's computer analysis of vocabulary in his assigned textbook. This study was undertaken to make more informed decisions for vocabulary teaching when using the textbook.

かつて、コンピューター言語学者の領域のみであったコーパス研究は、いまや基本のコンピューター技術さえあれば、素人でも簡単に手が届く範囲になっている。語彙の教え方や学習方法は利用可能であり、かつ具体的な教授法に応用できる研究領域である。この資料では、作者が指定した教科書におけるコンピューター語彙分析結果について述べ、その教科書を使用する際の、分析結果を用いた効果的な語彙の教え方を、さらに詳しく叙述している。

In many English courses in Japanese universities, teachers are required to use a commercially produced textbook, which often is the foundation for instruction in the course and thus students' principal source of vocabulary during the course. It may seem a straightforward business 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 many textbooks.

However, in order to make well-informed decisions about vocabulary instruction in their classes, teachers need to have the answers to some fundamental questions about the vocabulary content of their textbooks. These include "is the level of the vocabulary in this textbook appropriate for my learners?" The learner-level

of the textbook advised by the publisher can be a starting point in assessing the difficulty of a textbook's vocabulary, but it is a general indicator only, as publishers consider multiple language features in addition to the difficulty of the vocabulary, such as syntactic complexity and the length of both reading and listening texts. Teachers may also wonder "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 computer analysis of the vocabulary in the textbook being used, which can show how many and which words are in the first one or two thousand most frequent words of English, and which low-frequency words are present. It also shows if there too many of these words for the level of the learners, and how many times and in which units a word is recycled throughout the textbook.

The textbook

The textbook that was analyzed for this paper was *Firsthand Success: Beginners' Course 2, Gold Edition* (Helgesen, Brown, & Kahny, 2001), the assigned text for the author's yearlong university course. The textbook, which was written for high school and university use in Asia, is colorful and has an attractive design and an easily-navigated layout. Topics are typical of general-interest textbooks: Abilities, Jobs, Directions, and so on. It has an introductory unit, 10 principal units, 2 review units, writing units for each main unit, and learning summaries with key vocabulary and structures for each unit. The textbook is aimed at lower-level learners who have had some previous experience learning

English. In particular, the book aims to give students communicative practice to help activate the knowledge they may have acquired previously but are not able to use productively. Although it is presented as a four-skills textbook (each unit has a reading section and there are writing units, as mentioned), the skills focus is weighted more heavily towards listening and speaking. The textbook authors estimate a 40-50 hours course using the textbook.

Method

The textbook corpus

For this investigation, the main contents of the textbook (the units listed above) were converted into ASCII format text files, so that they could be "read" by the computer program used to analyze them. This was done by manually 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.

The material included in the corpus comprised what would typically be the core instructional material for the textbook, which most teachers could reasonably be expected to cover when using the textbook for a yearlong university English course. This included the contents of the textbook exclusive of acknowledgments, the introduction and other ancillary pages not part of usual classroom instruction, such as the

Resources page, which explains support materials such as the CD and website. 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.

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/downloads/NATION.zip>.

RANGE is a powerful yet easy-to-use program that analyzes a number of vocabulary features in multiple texts simultaneously. The contents of the texts are categorized into tokens, types, and word families. A token is a word in a text and a count of tokens is a count of the total number of words in a text, including repetitions of the same word. A type is a unique occurrence of a word: when counting types, a word is counted only once no matter how many times it occurs. 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*. For each type, *RANGE* gives a range or distribution figure (how many texts the type occurs in), the frequency of the type for each of the texts it occurs in, and a family frequency number (the total number of times the headword type and its family members occur in all of the texts).

The program also compares the contents of a corpus with word lists, to see which words are present in the 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* mentioned above) as well as their family members.

Results and discussion

Summary results

The data showing results after running *RANGE* over the textbook corpus appear in Table 1. The corpus consisted of 13058 tokens, comprised of 1407 types. 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 tokens were found in the list of the first 1000 words of English, and that these made up 82.2% of all tokens 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.

Table 1. Summary data of textbook corpus as analyzed by RANGE

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

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. As seen in Table 2, the coverage of the textbook by the first 2000 words was almost identical to the coverage of conversational language at 90.6% and 90.3%, respectively. The coverage was considerably greater – at least 8% higher – than any of the genres of fiction, newspapers, and academic text. This was an appropriate result, given that the focus of the textbook is on developing speaking and listening skills.

Suitability of the vocabulary level 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 across the lists and words not in the lists. We have seen that over 90% of the tokens in the textbook fell within the first 2000 words of English, a level frequently mentioned as the basic initial goal of many second language learners (Schmitt, 2000, p. 142); the content of the textbook would seem appropriately aimed

for students in the beginning stages of learning English. Furthermore, 82.2% of the textbook was covered by the first 1000 words, suggesting that most of the contents of the textbook are accessible to students whose word knowledge is in that range. This would make it a suitable text for low-level students who have had prior study experience studying English, the target group for this textbook. The author estimated his students to be mostly within the 400-600 words level, based on the level at which they were reading comfortably in their year-long extensive reading program, and the textbook would provide a significant but not overwhelming challenge for students in that general range.

Vocabulary load: Individual words

The numbers for coverage of the text by tokens provides a broad measure of the vocabulary load of the textbook but the number of types shows how much different vocabulary is present. This gives a clearer picture of the vocabulary load of the textbook and points the direction for supplementing the textbook vocabulary.

The summary data (Table 1) show a total of 1407 types, 798 of which appear in the first 1000 words. Yet this is not the nearly

Table 2. Coverage of different genres of English by word level

Word Level	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).

80% of the first 1000 words that it might seem. As mentioned earlier, *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, it is necessary also to look at how many word families from the first 1000 are represented in the text, in order to get a more accurate idea of how many of the first 1000 words of English are covered by the textbook. This number is 506; in other words, the textbook provides exposure to members of about 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 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 and would require substantial supplementation by the teacher.

Distribution and repetition of vocabulary and learning goals

Research shows that for a learner to have a realistic chance of learning a word it is necessary to encounter the word in a variety of contexts a number of times, at regular intervals (Nation, 2001, Schmitt, 2000). An analysis such as that provided by *RANGE* will show if target vocabulary occurs frequently enough and is given enough repetitions over time to provide suitable 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.

RANGE provides 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. 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.

(Table 3), and detailed results for all the words in the corpus, an excerpt of which appears in Appendix A.

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. 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 (tokens) of academic texts in between meetings of the same word. Nevertheless, their finding was for native-speaker academic 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, the detailed results of type distribution show the number of repetitions as well as the range of repetitions across the units of the textbook, data that allow the teacher to determine if a word is being suitably recycled in the textbook. Repetition of vocabulary items is important because students are unlikely to remember words that are encountered only once or twice. Students are aided in their vocabulary learning when words are repeated both multiple times and at regular intervals. Nation (2001) states that extensive memory research, including that specific to second language learning, shows that spaced repetition of vocabulary results in better retention rates than massed

repetition, a large amount of repetition in a limited time period such as multiple repetitions in one class period.

A 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—was chosen to illustrate how the detailed results reveal range and repetitions of words. 116 types occur in ten or more texts in the corpus and fifty-six of these, or 48%, are content words. The thirty-eight content words with the greatest range and frequency are shown in Appendix B, in a table that has been coded to show repetitions. It can be seen that these words have good range as well as a high rate of repetition. Almost all words appear in every main unit at least two or three times, which satisfies the conditions of spaced repetition. However, almost all these words have to do with classroom language and there are few instances of these words being used for anything other than instructions or section headings, with the exception of *like*.

In and of itself, this is not a negative result. Knowledge of these words is needed 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 during the course. Nevertheless, one would expect such words to be very well represented throughout a textbook. It is important also to look at words that are not used mainly for classroom language. A sampling of these appear in Appendix C, which shows all thirty of the content words with a range of six, in a table coded to show repetitions identically to the table in Appendix B. Few of these words

have primarily to do with classroom language. Rather, they are considered to be general service vocabulary, and represent typical vocabulary learning goals. As seen earlier, there are relatively few types overall in the textbook, so it is particularly desirable that the general service content words that are present 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*, and *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 and words not in the lists

Only 2.4% of types are found in the academic word list. This is unsurprising, as the textbook is not aimed at learners expecting to function in an English speaking academic environment. In order to decide if any of these are worth spending time with in class it is again necessary to look at the range and frequency of the items. None of the ten most frequent words 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*). 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 learning benefits, but the teacher would nonetheless want to check the list of words for any exceptions.

The most frequent types not in the lists are 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*), and words specific to the textbook context (e.g., the section headings *Duet* and *Solo*). These words make up 25% of types. This is not alarming because in most language contexts the majority of types fall outside the first 2000 words of English. However, the teacher may want to examine the words not in the lists to identify those needing attention in class.

Application of results

The author considered the results of the analysis when planning vocabulary instruction for the course. Since 82.2%

of the textbook was covered by the first 1000 words and only 8.4% was covered by the next 1000, a teacher could choose to focus on those words of a unit that appear in the first 1000 words, information available in the detailed results provided by *RANGE*. This is the course the author took, as he had limited time for explicit vocabulary instruction and study in class, and could not reasonably set targets beyond the first 1000 words. The author focused on the vocabulary that was in the textbook and did not supplement the textbook vocabulary with the words from the first 1000 words that did not appear in the textbook. With more time for vocabulary instruction, it would have been desirable to supplement the textbook to more fully cover the first 1000 words level. However, as only about half of the word families of the first 1000 words were present in the textbook, it would take considerable time to create supplemental materials and additional class time would be required for practice of the words. If the textbook covered the first 1000 words more completely, the words would be encountered in class and no supplemental materials or additional class time for practice would be necessary. Finally, because the range of most words was limited, the author provided supplementary practice for the vocabulary in each unit: students made vocabulary cards from lists compiled by the author and practice activities were used, including crosswords, gap fills and games using the vocabulary cards. Vocabulary quizzes were given for each unit. In this way, the author tried to give the students an improved chance of learning the vocabulary that was present in each unit.

Conclusion

One cannot reasonably expect a low-level, 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 to refresh the students' learning and give them needed vocabulary practice. Because the 2000 word level is an essential goal for second language learners, it is 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 examined here showed that the overall contents of the textbook reflected a difficulty level 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. However, the textbook does not contain adequate vocabulary items for a year-long course, and does not provide enough spaced repetition of the word items. Although it is comforting to think that because a textbook provides a list of important vocabulary for each unit the students will learn this vocabulary after using the textbook, this is not always the case. Even a limited analysis of range and repetition shows that it is necessary to extend the practice of vocabulary learning in the textbook if the students are to have a good opportunity to learn and retain it during the course. Furthermore, without significant supplementary materials provided by the teacher to extend vocabulary learning beyond the textbook, students have little chance of increasing their vocabulary to even the 1000 word level.

The analysis conducted in the paper was not intended to criticize a particular textbook. Indeed, the author has assumed that most textbooks suffer from similar problems as discussed in this paper. Is it possible to write a textbook that not only uses words mostly within the first 1000 or 2000 words, but also 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. However, while it entails considerable time and effort, an analysis such as the one presented here can help teachers to become aware of and compensate for shortcomings in the vocabulary content of their textbooks.

Mark Alberding teaches at the Center for English Language Education at Asia University in Tokyo. His interests include listening, vocabulary, and corpora and pedagogy. He can be contacted at markalberding@gmail.com .

References

- 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.
- West, M. (1953) *A General Service List of English Words*. London: Longman.

An earlier version of this paper appeared in *The CELE Journal* 13, 2005.

Appendix A: An excerpt of detailed distribution/range data for individual types

TYPE	RANGE	FREQ	U0	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	RV	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	
A	29	337	10	10	13	38	11	17	12	20	17	44	23	21	0	4	6	2	7	2	12	4	28	7	1	4	5	3	4	1	2	0	9	0	
TO	29	311	7	16	11	7	14	15	57	24	11	18	39	16	2	0	0	2	0	17	5	4	2	8	2	1	1	1	1	9	6	2	3	10	
USE	29	75	5	2	2	6	0	9	0	0	1	1	4	3	2	2	1	2	4	2	2	1	1	2	1	1	3	2	4	3	4	1	3	1	
THE	28	455	12	16	27	36	34	31	24	56	32	17	33	29	2	15	3	16	4	7	11	18	3	4	0	0	0	1	3	1	15	4	1	0	
I	27	350	5	13	10	9	25	23	71	15	47	18	16	9	3	0	5	11	5	8	0	4	0	0	4	1	0	6	8	19	3	8	2	2	
YOU	27	498	29	45	8	15	30	51	20	16	49	21	77	49	6	0	0	0	0	6	6	4	3	15	10	0	1	7	4	3	1	6	2	14	
IN	25	145	6	10	7	7	7	20	8	8	9	9	9	7	5	5	2	0	7	3	0	3	2	3	1	0	0	1	0	2	3	0	1	0	
AND	24	92	2	3	9	5	5	5	6	7	4	6	7	3	0	0	0	0	1	0	0	2	0	0	2	2	2	4	3	3	3	3	1	1	5
IS	23	205	0	22	20	23	4	3	4	9	2	17	6	19	5	5	14	12	0	0	6	0	17	0	5	5	1	0	0	0	2	1	3	0	0
WRITE	23	65	4	5	1	2	4	5	4	3	6	3	3	1	3	1	2	2	5	1	2	3	2	2	1	0	0	0	0	0	0	0	0	0	0
AT	22	92	4	3	2	3	3	1	2	6	12	4	8	12	1	0	0	4	0	0	7	0	0	6	1	1	0	1	0	0	5	2	0	4	
DO	22	236	20	34	5	5	20	20	16	7	2	11	20	25	7	0	0	0	0	8	0	0	0	1	11	0	0	8	3	5	1	0	1	6	
IT	22	113	3	5	12	5	3	13	4	13	5	3	13	0	0	0	0	0	0	0	0	0	3	0	1	4	6	1	2	1	5	2	3	6	
LANGUAGE	22	28	1	1	1	1	1	4	1	1	1	1	1	2	0	0	0	0	0	0	0	0	0	0	3	1	1	1	1	1	1	1	1	1	
S	22	121	5	5	14	8	3	6	1	15	3	12	10	11	0	6	0	0	0	0	0	0	0	0	3	4	4	1	1	0	6	1	1	1	
THAT	22	90	6	3	5	5	1	10	2	2	3	9	11	9	0	0	0	0	0	0	0	0	5	0	1	2	2	0	2	1	1	2	7	1	
ON	21	83	0	2	2	2	6	6	5	11	11	2	6	2	0	2	0	0	0	4	5	4	0	0	0	0	1	1	0	1	5	3	0	2	
WHAT	21	155	7	20	4	6	3	1	5	1	27	9	6	26	4	0	0	6	0	6	0	6	6	6	6	0	1	0	1	0	0	4	0	0	
WORDS	21	41	4	4	1	4	2	3	1	1	1	2	4	1	0	2	1	2	2	1	2	1	0	1	1	0	0	0	0	0	0	0	0	0	0
OR	20	87	2	1	6	5	27	3	2	6	1	1	4	6	0	0	5	0	0	0	0	0	1	0	0	2	1	8	0	0	1	0	3	2	
OUT	20	22	0	2	1	1	1	1	1	1	1	1	1	2	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	
T	20	178	9	0	2	4	3	30	13	2	7	7	45	9	0	0	0	0	9	1	0	0	0	11	0	1	0	0	4	4	0	2	2	13	
YOUR	20	183	3	28	12	10	18	11	13	12	18	13	15	2	0	1	0	0	0	0	0	0	1	4	7	0	0	0	1	0	0	0	0	0	1
ABOUT	19	62	1	4	7	4	7	3	5	3	6	6	4	1	0	0	0	0	2	0	0	2	0	1	1	1	0	1	0	0	0	3	0	0	
ARE	19	79	1	12	11	13	3	1	5	3	3	5	5	2	2	3	3	0	0	0	0	0	2	0	2	2	0	0	0	0	0	1	0	0	
LIKE	19	113	7	15	3	3	26	2	15	1	1	5	1	6	1	0	0	7	0	5	0	0	0	0	3	1	0	6	0	5	0	0	0	0	

Note: The table shows how RANGE displays the distribution or range of words across all the texts in the entire corpus. For the type in each row, the Range column shows in how many different texts the type occurred. The Freq column shows the overall frequency in the corpus of the type. Columns U0 through L10 (see Key below) show the frequency of a given type in a given text. For example, Can (bottom row) appears in 17 different texts a total of 171 times in those texts. It occurs four times in Unit 0 (U0), nine times in the two Review Units (RV) and not at all in Learning Summary 1 (L1). S, and T represent, respectively, either the possessive or a contraction of the copular be (the program cannot distinguish which), and the contraction of not.

Key: U= Main Unit, RV=Review Units 1 & 2, W=Writing Unit, L=Learning Summary

Appendix B: Content words with the greatest range

TYPE	RANGE	FREQ	U0	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	RV	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		
USE	29	75	5				6		9					4	3				4								3		4	3	4		3			
WRITE	23	65	4	5				4	5	4	3	6	3	3		3				5			3													
LANGUAGE	22	28						4																	3											
WORDS	21	41	4	4			4		3					4																						
LIKE	19	113	7	15	3	3	26		15			5		6				7		5				3			6		5							
EXAMPLE	18	30																																		
QUESTIONS	17	54		5	3	4	3	3	3	4	7	4	4	8																						
SAY	17	54	4					3	8		6		3	7	8																					
TIME	17	44		6					3	4		4		3	5	4																			3	
ENGLISH	16	56	10	7					9					9										4												3
	16	41							3		4	3			4				3	7															4	
READ	16	29	3						3	7				3																						
PEOPLE	15	44					6		6				13										3													
LISTEN	14	89		8	8	8	8	8	7	7	8	8	8	6																						
LOOK	14	28	3			3	4																													4
NOW	14	25												5	4																					
ANSWER	14	24										3			5																					
PAGE	13	35				3		3	3	3	4		3	3	4																					
FIND	13	34		5				4		3	4		4																							
THINK	12	43					4				3	3		5	10			5	6																	
GO	12	40		6					3	6	6		3		3											4										
ASK	12	32	4					5				5			6																					
BOOK	12	27						3	5					3																						
TAKE	12	23					3	3		3	3	3																								
=0DMUSIC	12	20																4																		
AGAIN	12	18																																		
MEAN	12	17	3												4																					
CIRCLE	12	16																																		
ACTIVITY	12	15		4																																
POINT	12	15							4																											
ONLY	12	14																																		
LEARNING	12	12																																		

Key: = 0 occurrences, = 1 occurrence, = 2 occurrences, U = Unit, RV= Review Units 1 & 2, W=Writing Unit, L= Learning Summary

Appendix C: Content Words From The Corpus With A Range of 6

TYPE	RANGE	FREQ	U0	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	RV	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		
FAVORITE	6	37		19	4	4								5	5										3											
STORE	6	28								17											4										6					
YES	6	20			6			8																	4											
PAST	6	18									8																						4			
SUMMER	6	16					5																					5								
HOPE	6	14							4						3						3															
ANIMALS	6	13				5														3																
CITY	6	13						3		4																										
ROCK	6	13						5																												
FRIEND	6	11									4														3											
SHOW	6	11			3													3																		
STUDY	6	11								3																										
=TRY	6	11								4																										
VISIT	6	11																																		
UNDERSTAND	6	10	5																																	
=ODFISH	6	9				3																														
NEW	6	9										3																								
SCIENCE	6	9					3																													
GAMES	6	8																																		
TALK	6	8												3																						
YEARS	6	8																																		
BIG	6	7																																		
EXAMPLES	6	7																																		
FAST	6	7																																		
FRIENDS	6	7																																		
GREAT	6	7																																		
MEET	6	7																																		
=ACTION	6	6																																		
=NAMES	6	6																																		
QUESTION	6	6																																		

Key: ■ = 0 occurrences, □ = 1 occurrence, ▒ = 2 occurrences, U = Unit, RV= Review Units 1 & 2, W=Writing Unit, L= Learning Summary