Evaluating the vocabulary load of written text

Stuart Webb and Paul Nation

Victoria University of Wellington

Teachers must choose many texts for students to read in language learning tasks. Teachers may select news articles, short stories, blog entries, and materials specifically written for students. Selecting texts that are at an appropriate level is essential for reading tasks to be effective. However, it can be extremely difficult for teachers to determine whether texts are at a suitable level for their students. Some texts may be understood while others are not. The coverage of a text (the percentage of known words) may indicate whether or not students are able to understand written input. The Vocabulary Size Test (Nation & Beglar, 2007) and the RANGE Programme (Nation & Heatley, 2002) can be used together to determine the coverage of a text and the words which are likely to be unknown. The aim of this article is to look at how teachers can use these tools to evaluate the appropriateness of texts.

Introduction

Laufer & Sim (1985) found that vocabulary knowledge may be the best gauge of whether or not a text will be understood. Research indicates that for learners to be able to guess words in context and gain adequate comprehension of written text it is necessary to know at least 95% of the words (Laufer 1989). Moreover, comprehension and incidental vocabulary learning through reading are likely to increase if the percentage of known words in a text is 98% (Nation, 2001). The vocabulary size necessary for reading comprehension may vary depending on the type of text and the degree of comprehension required. If 95% coverage is required, a vocabulary size of the most frequent 4000 word families may be necessary for comprehension of novels and newspapers. If 98% coverage is required, knowing 8000 word families may be necessary to understand newspapers, and knowing 9000 word families may be required to understand a novel (Nation, 2006). While these findings provide a useful guide for teachers, the question remains as to how do teachers evaluate the vocabulary loads of specific written texts used in the classroom? The vocabulary size necessary to understand individual texts varies from text to text and with shorter texts it may vary significantly. Estimating whether learners have the vocabulary size needed to reach 95% or 98% coverage of text is challenging and mistakes may lead to a lack of understanding for learners and difficulties in achieving the goals of the lesson for teachers. Two tools which teachers can use together to evaluate the vocabulary loads of text are the RANGE Programme and the Vocabulary Size Test.

The RANGE Programme

The RANGE Programme analyses the vocabulary in text. It allows the user to (a) determine the vocabulary size necessary to understand the vocabulary in text, (b) create word lists based on the frequency of occurrence and range of use of vocabulary in different types of discourse, (c) determine the number of encounters with words in a text, and (d) to evaluate the vocabulary load of text for teaching and learning language. The RANGE Programme is best known for the first three purposes. Most notably it was used in the development of the Academic Word List (Coxhead 2000) and the 14 British National Corpus (BNC) 1000 word lists (XXXX) as well as a series of studies by Nation and his colleagues which examined the vocabulary size necessary to understand different types of discourse.

Perhaps of most interest to teachers is the use of the RANGE Programme for analysing texts to be used in the classroom. Surprisingly its use as a tool for evaluating the vocabulary load of text has received little attention. This may be because originally the RANGE Programme was used together with West's (1953) General Service List and the Academic Word List to show how the most frequent 2000 words and academic vocabulary were represented in a text. Examining text using these lists allowed teachers to determine which words may or may not be known for learners with a relatively small vocabulary size. The development of the 14 BNC 1000 word lists for use with RANGE provides a more precise assessment of a text because it allows users to determine the vocabulary size necessary for comprehension of authentic texts which typically contain a small percentage of words ranging from the most frequent 3000-14000 word families.

The 14 BNC 1000 word lists were created according to the frequency and range of occurrence of word families in the BNC. The word families in the lists were created at Level 6 according to Bauer and Nation (1993) classification of word families. Level 6 word families include inflections and over 80 derivational affixes. All word stems were free forms not bound forms. Words which are not found in the most frequent 14000 word families may be classified as *proper nouns*, *marginal words*, and *Not in the lists* (items less frequent than the most frequent 14000 word-families). The RANGE Programme and the words lists are free to download from Paul Nation's website: http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx Using these lists with RANGE, teachers can measure text difficulty. However, for teachers to determine whether specific texts are suitable for specific learners, they also need to measure the learner's vocabulary size.

Teachers can use the software to analyse the vocabulary in a single text or as many as 32 texts at the same time. Analysing multiple texts at the beginning of a course may allow teachers to not only determine whether the texts are appropriate for the level of their learners but it also allows teachers to sequence the texts according to their vocabulary loads. Another useful feature of RANGE is that it can also show how many times each word is encountered in one text as well as in multiple texts. Because the number of encounters with unknown words may provide some indication of their potential for incidental learning (Webb, 2007) this may give teachers the opportunity to consider whether the learners may be able to learn unknown words on their own or whether items merit explicit attention.

The Vocabulary Size Test

The Vocabulary Size test provides an accurate and reliable assessment of learner vocabulary from the first 1000 (the most frequent 1000) to the 14th most frequent 1000 word families. It provides teachers with a more precise measurement of vocabulary size than the widely known Vocabulary Levels Test (Schmitt, Schmitt, & Clapham 2001) because it measures knowledge of vocabulary size at 14 different points, and all of the points are at an equal distance apart (1000 word families). It is slightly more demanding than the Vocabulary Levels Test because greater knowledge of the meaning of the items may be necessary to answer questions. The Vocabulary Size Test measures learners' knowledge of the same 14 1000 word lists that can be used with RANGE. Because both RANGE and the Vocabulary Size Test involve the same 1000 word lists, using them together may be the most effective method of determining which words are likely to be known and unknown for specific learners.

The Vocabulary Size test uses a multiple choice format with ten questions measuring knowledge of each 1000 word level. Because there are ten questions per 1000 word level, each question represents knowledge of 100 word families. This means that if learners had a score of 9/10 on a particular level, then they would have demonstrated knowledge of 900/1000 word families from that level. The test is easy to administer and takes little time to grade. The following are two questions taken from the first 1000 word level:

TIME: They have a lot of **time**.

a. money
b. food
c. hours
d. friends

JUMP: She tried to **jump**.

a. lie on top of the water
b. get off the ground suddenly
c. stop the car at the edge of the road
d. move very fast

It is important to note that the Vocabulary Size Test measures knowledge of word families. Thus, if learners are able to demonstrate knowledge of the headword in the test, there is an assumption that they also have receptive knowledge of the rest of that word family. The following are examples of the word families for *time* and *jump* in the first 1000 word level:

TIME	JUMP
TIMER	JUMPS
TIMES	JUMPING
TIMELESS	JUMPED
TIMELY	JUMPY
TIMING	
TIMED	
UNTIMED	

Procedure for evaluating text

The following sections show how the Vocabulary Size Test and the RANGE Programme can be used to evaluate a text.

1. Measure the vocabulary size of the learners

The best time to measure vocabulary size is at the beginning of a course because it will provide useful information on whether or not learners have the vocabulary necessary to do certain tasks, and it may also allow teachers to chart learners' vocabulary development through a language learning program. Because the Vocabulary Size Test contains 14 tests, each measuring knowledge of 1000 word families, teachers may not wish to administer all of the sections of the test if their students are at a beginning or intermediate level. It may be sufficient to administer the first five sections measuring knowledge of the most frequent 5000 words for less

advanced learners. For the purposes of this article, the following test scores will be used as an example:

Vocabulary Size Test scores

Test	Mean score
1000	9.7
2000	9.1
3000	8.5
4000	4.8
5000	1.3

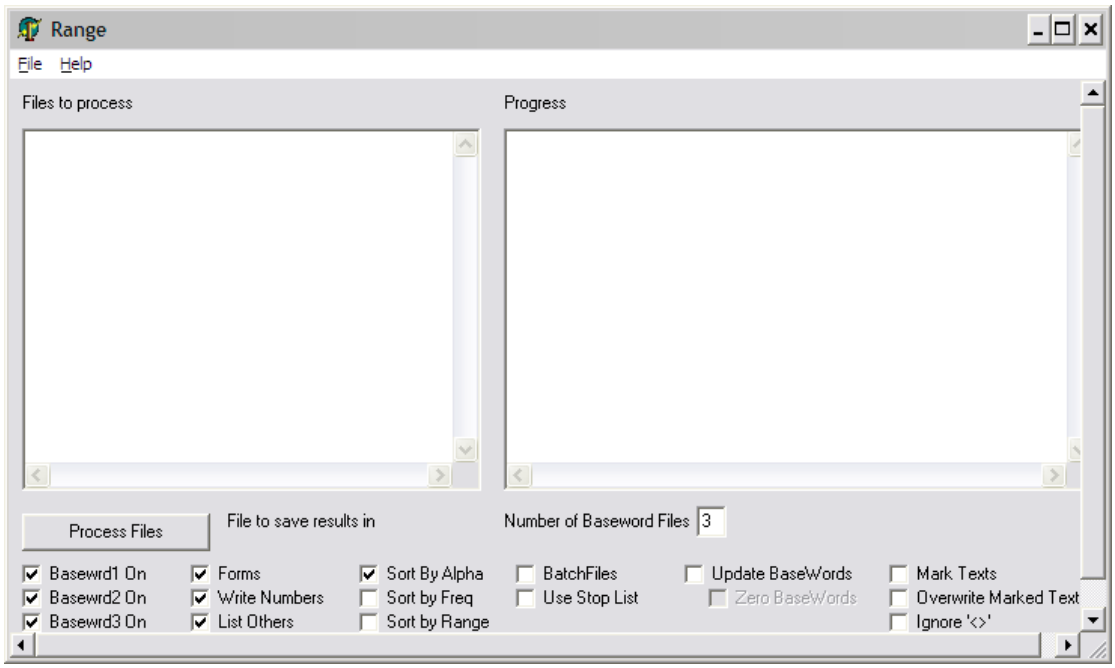
The example above indicates that as a whole the learners' vocabulary knowledge is limited to the most frequent 3000 word families. However, teachers should examine individual tests to determine whether some learners may still need to learn some of the third 1000 word list. Overall, the test profile indicates that teachers can assume knowledge of the most frequent 3000 words for the majority of students. The mean score on the fourth 1000 word test indicates that most learners still need to work on learning words from this list. Teachers should therefore assume that words which are less frequent than 3000 word families might be unknown.

2. Analyse text using the RANGE Programme

Once the vocabulary size of the learners is known, the RANGE Programme can be used to determine the vocabulary size necessary to reach 95% and 98% coverage of text, as well as the number of words in the text which are likely to be unknown. The

newspaper article *Canada's 2nd fastest supercomputer assigned to massive cancer project* was selected from the online version of *The Toronto Star*, because its length made it suitable for use in a variety of different types of classroom tasks. Once teachers have selected a text, it is a simple process to analyse it. First, teachers need to convert the text into a text file. Next, open the text using RANGE and indicate the text file in which the output is to be saved. The *Number of Baseword Files* should be changed from 3 to 16 for use with the 14 BNC word lists. The 15th and 16th lists show the proper nouns (List 15) and the marginal words such as ah, oh, huh (List 16) which occurred in the text. Finally, clicking on the *Process Files* button will complete the analysis. The RANGE Programme interface is shown in Figure 1. The time taken for RANGE to analyse the text depends on the number of tokens. Range completed the analysis of the text analysed in this article in 1 second. Teachers can then open the file where the output has been saved to see the analysis. More detailed instructions and a troubleshooting guide come with the software and word lists.

Figure 1. The RANGE Programme interface



The RANGE Programme output provides two valuable tools for teachers. The first is an analysis of the text's distribution of tokens, types, and word families in each of the different 1000 word lists. It is useful for teachers to understand the difference between the three categories. Each word in a text counts as one token and each different word in a text counts as one type. Thus, the previous sentence consists of 20 tokens and 12 types because 8 of the words in the sentence occur twice. Examples of word families were given above. The second part of the output which is useful is the data for each word found in the text. This shows where each word in the text was categorized by the word lists as well as the number of times it was encountered in the text. The RANGE output which shows the distribution of the various levels of vocabulary in the text was as follows:

WORD LIST	TOKENS/%	TYPES/%	FAMILIES
one	243/69.23	109/62.29	94
two	27/ 7.69	19/10.86	17
three	24/ 6.84	15/ 8.57	12
four	11/ 3.13	7/ 4.00	7
five	5/ 1.42	3/ 1.71	3
six	4/ 1.14	3/ 1.71	2
seven	3/ 0.85	2/ 1.14	2
eight	2/ 0.57	1/ 0.57	1
nine	0/ 0.00	0/ 0.00	0
ten	0/ 0.00	0/ 0.00	0
11	1/ 0.28	1/ 0.57	1
12	0/ 0.00	0/ 0.00	0
13	14/ 3.99	3/ 1.71	2
14	1/ 0.28	1/ 0.57	1
15	6/ 1.71	4/ 2.29	4

16	0/ 0.00	0/ 0.00	0
Not in the lists	10/ 2.85	7/ 4.00	?????
Total	351	175	146

The first column of the output shows the different 1000 word lists, and the second, third and fourth columns show the number and percentage of tokens, types and word families according to each 1000 word list. In the final row of the output we can see that there were 351 tokens, 175 types, and 146 word families in the text. The output clearly shows the relative importance of knowing the most frequent words. Over 69% of the words (243 tokens) were from the 1000 word list, 7.69% of the words (27 tokens) were from the 2000 word list, and 6.84% of the words (24 tokens) were from the 3000 word list. Thus, if the most frequent 3000 word families are known to readers, they would know 294 of the 351 tokens in the text. Many researchers have taken the approach that proper nouns have a minimal learning burden and may be easily understood by readers. If we add the proper noun tokens in List 15 to the total number of tokens that would be known for readers who know the most frequent 3000 words then 300 of the tokens would be known leaving 51 unknown words. If the percentage of tokens known to the learners reached 95% coverage, the text would be suitable for classroom use without any assistance from the teacher. However, in this case only 85.4% of the tokens would be known to learners with a vocabulary size of approximately 3000 words. This indicates that the text would be too difficult for the learners to read without assistance and that further analysis is needed to determine if the text may be suitable for use.

It is also useful for teachers to look at the results in terms of word families because similarity in forms and meanings for tokens from the same family may reduce the difficulty of learning word families. Teachers can cross-reference tokens with word families by looking at the same row of columns two and four. For example, the second column of the output shows that there were 14 tokens which occurred in the 13th 1000 word list. However, when we look at the same row in the word families column we can see that those 14 tokens consist of only two word families. If we count the number of word families which may be unknown for learners with a vocabulary size of the most frequent 3000 words, we find that there are 19 different word families plus a maximum of 10 word families for the 10 tokens classified as 'Not in the Lists.' Therefore, if the scores on the Vocabulary Size Test indicated that the learners had an average vocabulary size of 3000 word families, there would be a maximum of 29 unknown word families in the text. We can see each of these word families if we scroll down in the RANGE output. The RANGE Programme lists the word families according to the 1000 word lists so it is easy to find the words which may be unknown.

The following output from RANGE shows the different word families that occurred in the text listed by their relative word frequency in the BNC. The column labelled RANGE shows the number of texts in which the word occurred. In this example, there was only one text analysed so every word was found in only one text. If multiple texts are analysed together, it is very useful for teachers to look in this column to see the number of texts in which each word occurred because further encounters with recently encountered/learned words may deepen knowledge of those words. The TYFREQ and FAFREQ columns show the number of occurrences of the exact word type and word

families that occurred in the text. For example, at the 4th 1000 level (Base Four) the exact word type *acquire* did not occur, but one or more of the family members did. Whereas, we can see that the word family with the headword *grid* in the 4th 1000 word list occurred three times in the text, and each time it occurred as the word type *grid*. The output clearly shows which word families may be unknown or unfamiliar to learners. Teachers can quickly scan these words to determine if any of them may be known. For example, the compound word *supercomputer* may be unfamiliar to readers but it may be easily understood if the high frequency words *super* and *computer* are known. It occurs 11 times in the text so comprehension of this word may play a key role in understanding the text.

BASE FOUR FAMILIES	**RANGE**	**TYFREQ**	**FAFREQ**
ACQUIRE	1	0	2
GRID	1	3	3
INTERACT	1	0	1
MINISTRY	1	1	1
NETWORK	1	2	2
PROFILE	1	0	1
TASK	1	1	1
BASE FIVE FAMILIES	**RANGE**	**TYFREQ**	**FAFRE**
CLUSTER	1	2	2
COMPATIBLE	1	1	1
INNOVATE	1	0	2
BASE SIX FAMILIES	**RANGE**	**TYFREQ**	**FAFRE**
PROTEIN	1	1	3
SNAIL	1	1	1
BASE SEVEN FAMILIES	**RANGE**	**TYFREQ**	**FAFRE**
BUFFALO	1	2	2
CLASSIFICATION	1	1	1
BASE EIGHT FAMILIES	**RANGE**	**TYFREQ**	**FAFRE**
PC	1	2	2

BASE 11FAMILIES	RANGE	TYFREQ	FAFRE
TRILLION	1	1	1
BASE 13FAMILIES	**RANGE**	**TYFREQ**	**FAFRE**
IBM	1	3	3
SUPERCOMPUTER	1	10	11
BASE 14FAMILIES	**RANGE**	**TYFREQ**	**FAFRE**
CPU	1	1	1
BASE 15FAMILIES	**RANGE**	**TYFREQ**	**FAFRE**
IGOR	1	1	1
MARGARET	1	1	1
ND	1	1	1
ONTARIO	1	3	3

Types Not Found In Any List

TYPE	RANGE	FREQ
COMPUTER-GENERATED	1	1
CRYSTALLOGRAPHY	1	1
DISEASE-RELATED	1	1
HAUPTMAN-WOODWARD	1	1
HIGH-RESOLUTION	1	2
IN-KIND	1	1
JURISICA	1	3

In the output, we can see that there are also two more proper nouns from the text (the city of Buffalo and IBM) which are included in the 1000 word lists. These words along with *supercomputer* may be easily understood for learners with a vocabulary size of the most frequent 3000 words. The final section of the output *Types Not Found In Any List* shows the word types that were not sorted into any of the 16 word lists. The output in *Types Not Found In Any List* should contain the lowest frequency words found in the text. However, it may list a small number of known words because it lists

hyphenated words and some proper nouns which are not found in the proper nouns list. In the example, there are two proper nouns (Hauptman-Woodward and Jurisica), and four hyphenated multi-word items which contain words from the first and second 1000 word lists. Multi-item units may cause problems depending on the overlap in meaning between the individual words and the multi-word item. In the example, *computer-generated*, and *disease-related* may be understood for learners who know each of the individual words. However, *high-resolution* may be unknown to learners because *resolution* occurs in the 4000 word list. *In-kind* may also cause greater difficulty because there is little overlap between the individual words and the multi-word unit. *Crystallography* is the only individual word which is low frequency in the list. If we subtract the items found in the output above which may be understood by the learners, that leaves 19 unknown word families and 29 tokens in the text. This represents 91.7% coverage indicating that comprehension of the text would still be difficult without some form of learning support for the learners. Teachers must then consider whether the number of unknown tokens and word families would allow them to pre-teach or modify the text to make it appropriate for learners. The following section briefly describes options for increasing comprehension if texts do not reach 95% coverage.

3. Evaluate the potential use of the text for the learners

Using the results of the Vocabulary Size Test together with the RANGE output will leave teachers with several options depending upon whether there is a gap between the learner's vocabulary size and the vocabulary size necessary to reach 95% or 98% coverage. If the learner's vocabulary size is sufficient to reach the desired coverage

then there may be no need to assist learners with comprehension of the text. However, if there is a gap then it may be necessary for teachers to do one of the following:

1. pre-teach vocabulary,
2. ensure that dictionaries are available and learners are given time to look up unknown words during the reading task,
3. provide glosses for unknown words,
4. simplify the text
5. eliminate that text for use in the classroom and select a more appropriate text.

If the target coverage can be reached through learning a small number of unknown word families, teachers may wish to pre-teach those items or have students use dictionaries when reading the text. Both pre-teaching and dictionary use have been found to facilitate comprehension (see for example, Nation, 2001). However, it is important that there are a manageable number of words for learners to learn and look up. If there is a larger number of unknown words, glossing may be more useful. Glossing allows learners to read a text relatively quickly by checking the relevant definitions or L1 translations when necessary. Another option when there are many unknown word families would be to simplify the text by replacing the unknown words with known words or phrases which are likely to be understood by learners. This will involve more effort from teachers and may not be a valid solution for longer texts with a large number of unknown words. Because RANGE is fast and easy to use, simply eliminating the text from consideration may often be the best solution if learners do not have the vocabulary size necessary to understand the words in the text. Selecting alternate texts and analysing them with RANGE together with knowledge of

the Vocabulary Size Test scores may allow teachers to find more suitable texts for classroom use in a relatively short time. In the example above, for learners who knew the most frequent 3000 word families, the small number of unknown word families remaining would make pre-teaching or looking up words in dictionaries a viable option along with glossing or simplifying. If learners had a vocabulary size of the most frequent 2000 word families, there would be 12 more unknown word families or 24 more unknown tokens. This would make pre-teaching and dictionary use much more demanding. In this case, glossing and simplifying may be more appropriate. If learners only knew the most frequent 1000 word families, this text should probably be eliminated for consideration for use in the classroom.

Determining whether texts are appropriate for specific learners can be a difficult task for teachers. When the vocabulary is too difficult learners may not understand texts leading to problems for both teachers and students. Teachers can use the RANGE Programme together with the Vocabulary Size Test to quickly evaluate the vocabulary load of texts. Although researchers may be quite familiar with these tools, they may not be well known to teachers. The aim of this article was to demonstrate that evaluating texts with these tools can be fast, easy, and effective.

References

Bauer, L., & Nation, I.S.P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253–279.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34, 213-238.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special Language: From Humans Thinking To Thinking Machines*, (pp 316-323). Clevedon: Multilingual Matters.

Laufer, B. & Sim. D.D. (1985). An attempt to measure the threshold of competence for reading comprehension. *Foreign Language Annals*, 18, 405-411.

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

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

Nation, I.S.P & Beglar, D. (2007). A vocabulary size test. The Language Teacher, 31, 9-13.

Nation, I.S.P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts [software]. Downloadable from http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx

Schmitt, N., Schmitt, D. & Clapham. C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing* 18, 55-88.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28, 46-65.

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