New Zealand Studies in Applied Linguistics, 2010, 16 (2) 37-52

WHY MIGHT SECONDARY SCIENCE TEXTBOOKS BE DIFFICULT TO READ?

Averil Coxhead, Liesje Stevens & Jenna Tinkle

Victoria University of Wellington

Abstract

This study focuses on the vocabulary in four high school level Science textbooks in the Pathway Series published by New House in Auckland for years nine through twelve (Hook, 2004; 2005; 2006; Relph, Croucher, & Castle, 2006). The target audience for this series is high school students aged between 12 and 17 years. The study investigates the coverage of West's (1953) General Service List of English words *(GSL), Coxhead's (2000) Academic Word List (AWL), Coxhead and Hirsh's (2007) pilot science list for EAP over the textbooks. It also uses Nation's (2006) frequency lists, based on the British National Corpus, to find out how many words are needed to read these textbooks. Finally, we discuss the implications of this study for teachers and discuss how teachers might investigate the vocabulary in their textbooks and other classroom materials.*

Introduction

The impetus for this study of the vocabulary load of a series of secondary school textbooks came from conversations with several high school teachers. These teachers were asking about the efficacy of teaching, for example, words in the Academic Word List (AWL) (Coxhead, 2000) to English as an Additional Language (EAL) secondary school students in Aotearoa/New Zealand, as a way to support their learners with reading. The teachers were concerned that classroom textbooks were too difficult for their students to read. These teachers are not the only ones who are concerned about the vocabulary in textbooks, whether for non-native readers of English (see Brown, 2007) or native readers of English (see Fang, 2006, for example). The reciprocal relationship between vocabulary knowledge and textbooks is critical. Biemiller (2001) shows that vocabulary deficiency in early schooling tends to be exacerbated throughout later schooling. That is, gaps in vocabulary knowledge can widen as children get older. Knowledge of words alone does not guarantee that students will be better able to read their textbooks. Learners need instruction that encourages them to explore the relationships between words and key concepts, and encounter and use these words meaningfully to increase their ability to understand subject area texts (Harmon et al., 2000, p. 270).

Another separate discussion is whether it should be the teachers who bear the responsibility for teaching students how to read subject-specific books (Chall &

Conrad, 1991) or whether publishers should take responsibility for ensuring the content and vocabulary of their books are comprehensible.

A key concept in this study is the issue of who is a native speaker and who is not a native speaker. In New Zealand, secondary school classrooms contain a mix of students who speak languages other than English at home. Some students were born in the country or were born elsewhere and arrived after the age of five. In English-medium schools, these students should develop a native speaker vocabulary of English. For those who arrive after the age of eight or nine, a vocabulary test is needed to assess their vocabulary knowledge (see Nation, 2001 for explanations and examples of such tests). The main problem to consider in this area is that, if some students in a secondary class already have a vocabulary of up to 9,000 words, including Coxhead's (2000) Academic Word List (AWL), then we need to aware that we might be teaching students words they already know (Paul Nation, personal communication, 10 September, 2010).

Textbooks in secondary schools

Textbooks play a key role in educational delivery at all levels according to Harmon et al. (2000) in their analysis of vocabulary instruction in social studies textbooks from grades four to eight in the USA. However, this role is not without challenges. One problem for EAL students is understanding textbooks and another is a lack of support for learners' reading from publishers (Harmon et al., 2000, p. 253). In informal discussions with teachers, we found that comprehensibility was an important factor in choosing a textbook to use in class. Teachers also considered how much students enjoy a textbook, the fit of a textbook with the New Zealand Curriculum, and its fit with assessment criteria such as specific National Certificate of Educational Achievement (NCEA) standards. Textbooks are not the only materials in a classroom, however, as teachers may draw on web pages, YouTube, Smartboard, and other in-house materials. These changes in focus are reflected in a survey on book spending (Bray & Liew, 2008) which shows that secondary school spending on textbooks for classrooms has decreased by 40% between 1989 and 2008. In contrast, spending on IT has increased 293% since 1997.

Early studies of textbooks emphasised measuring readability, for example, by looking at sentence length or words per sentence or the number of syllables per sentence (see Chall & Conrad, 1991). Taylor's (1979) corpus-based study of secondary textbooks in Australia (1979) looked at textbooks across the curriculum areas of maths, science, history, commerce, social studies, and geography. One of Taylor's findings is that science books are more difficult to understand in terms of vocabulary than the other subject areas.

Vocabulary size and reading

Vocabulary-based studies have looked at how many words are needed to read a range of texts. Nation (2006) finds that 8,000-9,000 word families are necessary to read a range of authentic texts such as novels and newspapers. These estimates stand in stark contrast to Nurweni and Read's (1999) finding that students learning English in Indonesia have a working vocabulary of 2,000-3,000 words after years of high school instruction in their country. The key point here is that the bigger a learner's vocabulary, the more efficiently the learner can read. The more the learner reads, the more encounters the learner has with vocabulary in context. Researchers have sought to identify the most useful words for vocabulary learning as a means to help to close the gap between learner vocabulary size and vocabulary needs. That is, what words will readers encounter more often in a range of texts?

Vocabulary lists

In this section, we look at four existing word lists that are used in this study to investigate how many words are needed to read secondary school science textbooks. It is important to consider the development and purposes of these word lists.

West's *General Service List of English Words*

West (1953) developed the *General Service List of English Words* (GSL) using several principles for selecting words. One of the principles, for example, was the coverage of a word. This principle meant that West selected one word over a synonym, for example *work* instead of *job*, because one of the words 'covered' more than the other. For this reason, *end* is in the GSL whereas *final* is not. The list is well known and well used but it is dated. Another related problem is that it does not contain commonly used words that have entered the lexicon since it was developed (*computer*, *television*, *internet*, and *email*, for example) (Richards, 1974, p.70). However, West's GSL performs well over a range of texts, by covering roughly 90% of fiction texts (Hirsh, 1993), 75% of nonfiction texts (Hwang, 1989), and 76% of a corpus of written academic English (Coxhead, 1998).

As yet, no replacement for the GSL has been developed. However, we will see below how Nation's BNC lists present different frequency-based sets of up to 20,000 words that allow us the opportunity to look at a wider range of vocabulary within spoken and written texts. West's *General Service List of English Words* remains a remarkably durable and principled word list.

The Academic Word List

Coxhead's Academic Word List (AWL) (2000) was developed for EAP teachers to be used as part of their programme to prepare learners for university study or to be used by students working independently towards that goal. Coxhead excluded the 2,000 words of West's *General Service List of English Words* (1953) from the Academic Word List written academic corpus of 3.5 million running words. The corpus contained 28 subject areas within four academic disciplines. The resulting AWL contains 570 word families. The coverage of the AWL over the corpus is 10%, and over each of the four disciplines coverage is: arts 9.3%, commerce 12%, law 9.4%, and science 9.1%. Coxhead's analysis included coverage of a fiction corpus (1.4%) and a newspaper corpus (roughly 4%). These figures indicate this list is clearly academic in nature. Other studies have shown similar coverage figures over a variety of written academic corpora, see for example, medical research articles (Chen & Ge, 2007) and agricultural science (Martínez, Beck & Panza, 2009). The AWL is used by teachers and learners in many English for Academic Purposes classrooms around the world. See Coxhead (n.d.) for more information about this list.

A study of science-specific texts at middle school in the USA was carried out using the AWL. Greene (2008) compiled an 18-million word corpus (109 texts) of middle school textbooks. Middle school students in the USA are usually aged between 11 and 14 years old. Greene's corpus contained textbooks from five subject areas: English, health, math, science, and social sciences/history. Her analysis shows that while the AWL occurs in these textbooks, it does not occur as often as it does in university level texts (5.98% in a middle school science text). This finding indicates that middle school textbooks tend to be written in a relatively reader-friendly way and thus are not as scholarly as tertiary texts. Greene developed specialized middle school word lists for classroom teachers and materials design because there was no list to cover the middle school content areas adequately. The science section of the Greene corpus, for example, contained 3.7 million words. The science-specific word list from this corpus contained 3,028 words. When the GSL, AWL (5.98%), and Middle School (11.95%) lists are combined, the coverage over the science subcorpus is 89.18%. Unfortunately, the Greene lists are currently unavailable (Jennifer Greene, personal communication, 22 September, 2009). However an online webinar by Coxhead and Greene (2009) contains data and examples from Greene's study.

A pilot science-specific word list

One question about specialisation of vocabulary is where a more general word list might stop and a more specialised word list might start. That is, when is it more beneficial for learners to focus on more specialised vocabulary in a particular area? Coxhead and Hirsh (2007) addressed this question through an analysis of the GSL and AWL in a 1,761,380-word corpus of 14 tertiary science subjects and developed a pilot science-specific word list. The researchers found their resulting science list of 318 word families covered 3.79% of their written science corpus. The coverage of this science list was also well under 1% for each of the arts, commerce and law sections of a written academic corpus compiled by Coxhead (2000) and 0.27% of a 3.5 million word corpus of fiction. Coxhead and Hirsh (2007) note the principles for the selection of words in the GSL and AWL have an impact on the words in their science-specific list.

Nation's British National Corpus (BNC) Lists

A recent innovation in word list development is Paul Nation's work using the British National Corpus (BNC). Nation uses the spoken part of the corpus to develop these lists. He argues that people learn to speak almost without exception, while written language is not necessarily developed in the same way as spoken language (see Nation, 2004). Therefore spoken English data from the corpus reflects people's knowledge of vocabulary. Nation (2006, p. 79) points out that the amount of vocabulary knowledge needed for listening and reading are different in that,

> If we take 98% as the ideal coverage, a 8,000–9,000 word-family vocabulary is needed for dealing with written text, and 6,000–7,000 families for dealing with spoken text.

Nation (2006) explores the number of words needed to reach 98% in a variety of texts using the BNC lists. This research is closely linked to different purposes for learning language. For example, 8,000 to 9,000 words are for reading a novel, 3,000 words for reading a graded reader, and 7,000 words and proper nouns for watching the movie *Shrek*. See Webb and Rodgers (2009) on the vocabulary in movies and Webb (2010) for a discussion on the use of glossaries to increase vocabulary coverage of two well-known television programmes. The present study builds on Nation's findings by looking at secondary science textbooks and finding whether similar levels of vocabulary are needed for reading these texts.

To the best of our knowledge, no study has looked into the vocabulary of science textbooks in New Zealand. For the present study, the corpus of science textbooks will be investigated using existing word lists: the GSL, AWL and science-specific lists. The purpose of this part of the study is to investigate whether these lists contain words that are also in the textbooks. If they do, then they can be used with some confidence to guide teachers in their selection of words to teach in classes. The second list is Nation's British National Corpus frequency-based lists. This second measure is used to provide a view of the vocabulary that is needed to read the textbooks. These lists rank the words by frequency.

Research questions

We have two research questions for this study:

1. What coverage do the GSL, AWL, and science-specific lists have over the science textbook series?
2. How many words do learners need to know to read secondary science textbooks?

Methodology

Our study began by considering how we might investigate the vocabulary of specific subjects in secondary schools. We collected several textbooks in history, geography, and science and ran a small pilot study to test out our scanning capabilities. We found that the history and geography textbooks were difficult to scan accurately because of formatting and contrast problems. The science textbooks in our pilot, on the other hand, were more easily scanned. We found a series that is readily available from libraries and, according to the publisher is a bestseller in New Zealand. We scanned all four textbooks from the *New Zealand Pathfinder compact course book series* of Science textbooks for years nine, ten, eleven and twelve (Hook, 2004; 2005; 2006; Relph, Croucher, & Castle, 2006) and converted them to text using MS Office Document Imaging with OCR text recognition.

The books are colourful and set out with many pictures and textboxes and include many scientific symbols. We encountered two specific problems when scanning them. First, the complex layout of pages meant that a linear representation of the text was not achievable. That is, the scanned text did not always follow the same pattern as the original text because text was scanned across the page regardless of columns or textboxes. This problem caused us to check each page of the scanned text to ensure we had obtained as much of the text as possible. This task was made trickier because we could not always match the source and scanned texts. The second problem was sections of text with lighter colours such as pale yellow with white did not scan well. This problem meant that the text recognition process was sometimes not accurate. Our guiding principle was that we would correct errors in the scanned texts and check them against the corresponding text in the original. We also checked words where the word was recognisable and contextually correct. For example, if the word *thermometer* did not scan well and appeared as *thermomxter**, we corrected the word. Finally, we found and replaced hyphens with spaces in the corpus so that the words making up the hyphenated words were counted as single lexical items by Range (Heatley, Nation & Coxhead, 2003), the computer programme we used to analyse the texts.

The total number of running words in the corpus of textbooks is 279,733. The year twelve textbook is the shortest with 56,058 running words while year eleven is the longest with 88,685 running words. This imbalance between the numbers of words in each textbook presents difficulties for comparison because longer texts provide more opportunities for words to occur and recur. Table One below shows the total of tokens (or words) in each textbook.

Table One: Running words of the textbooks

	Year 9	Year 10	Year 11	Year 12
Running words	62, 757	72, 233	88, 685	56, 058

Analysing the data

The textbooks were analysed using the Range Programme (Heatley et al., 2003; see also Webb & Nation, 2009 for a description of this tool). Two versions of Range were used. One contains base files of the first and second thousand words of the GSL (headwords and family members), the AWL, and the science-specific word list. The second version uses the first 20,000 word families of the British National Corpus. For the present analysis, 23 files of word lists were used, from the first 1,000 up to 20,000 words. File 21 is a collection of proper nouns, which Nation is adding to as he develops his lists. File 22 is a small list of hesitation devices used in speaking, such as *um* and *ah*, which reflects the spoken origin of the lists. This data is not reported on here because it is not relevant to our study of written English. File 23 is transparent compounds, such as *arrowhead* and *breathtaking*. The final file contains a list of proper nouns. These lists and the Range programme are available from Nation's website (http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx). For a discussion of the 20,000 BNC word lists, see Nation and Webb (2010). Analysing the texts using different base word lists gives us a different view of the vocabulary in these textbooks.

Results and Discussion

Research question one: What coverage do the GSL, AWL, and science-specific lists have over the science textbook series?

To answer this question, we carried out an analysis of the corpus using the Range program (Heatley et al., 2003) using the GSL, AWL, and science-specific lists. We included Nation's proper noun list in this analysis because it is useful to see these figures alongside the data from the word lists. The results in Table Two below show that 89.1% of the words in the textbooks also occur in these four word lists, plus 0.48% coverage from 269 word families in the proper noun list, which makes 90.39%. The four word lists plus proper nouns contain a total of 2874 word families. Of these families, approximately 76% (2207) appear in the textbooks.

We can break that down further into 88.7% of the words in the first 1,000 list, 65% of the second 1,000 list, 72% of the AWL, and 83% of the science list.

Table Two: Coverage of the four science textbooks by the GSL, AWL, and Science-specific word lists

Word list	Tokens/%	Running coverage	Families
GSL 1000	70.10	70.10	886
GSL 2000	6.86	76.96	645
AWL	7.05	84.01	412
Science-specific list	5.90	89.91	264
Proper nouns	0.48	90.39	269
not in the lists	9.61	100	????

It seems that a number of the families in the second 1,000 words of the GSL are working hard in these textbooks. It could be the case that the second thousand words of the GSL contains more of a 'high school' vocabulary. Table Two indicates that the coverage of the first 2,000 words of the GSL is 76.96%, nearly 6% higher than the 71% coverage of the list over the tertiary science corpus from Coxhead and Hirsh (2007). This higher figure indicates that the textbooks contain more words from the GSL than the tertiary corpus does. Secondly, the coverage of the second 1,000 words of the GSL at 6.86% is higher than the coverage over the AWL corpus as a whole (4.7%) and the science-only section of that corpus (5%). One of the reasons for this higher coverage of the second thousand of West's GSL in these science textbooks is that ten words from the list occur over 250 times. These words are *electric*, *ray*, *temperature*, *solid*, *liquid*, *solve*, *false*, *reflect*, *angle*, and *plate*. These words are closely related to topics in science and to particular areas in science, such as *plate* in geology. The occurrence of these words is not uniform across all the textbooks. For example, *plate* occurs in texts for years ten through twelve, but not for year nine. This is because the former textbooks contain chapters on earthquakes whereas the latter does not. Other examples of uneven occurrences of these words include *ray*, *liquid*, *reflect*, and *angle*.

The AWL covers 7.05% of the textbooks. This figure is 2% lower than the coverage of the AWL over the science section of the AWL corpus at 9.1%. A total of 158 word families in the AWL do not occur in textbooks, which roughly equates to 27% of the words in the AWL. This lower coverage figure supports Greene's (2008) finding (and the secondary school teachers' observation) that the secondary school textbooks do contain AWL words but not to the same extent as university level texts do.

The science list by Coxhead and Hirsh covers 5.9% of the textbook corpus. This figure is higher than the coverage the researchers found over their tertiary science corpus (3.79%). This coverage figure suggests that the pilot science list contains a large number of words that are used in both secondary and tertiary level texts. A total of 54 of the word families in the list do not appear in this corpus, which is roughly 17% of the words in the list. A reason for this difference might be that the tertiary corpus is roughly six times the size of the textbook corpus.

Below is an example of words from these four lists in a section of one of these textbooks. The sample of text below is from Hook's (2005) year nine textbook, from the chapter on atoms (p. 88). In this sample, the words from the GSL are in regular text, the AWL words are in **bold**, the science list words are shaded, and the words not found in any list are in *italics*.

> Model B: *Mini* Solar System
>
> As scientists **investigated** atoms more closely they discovered that atoms were actually made of smaller *sub* atomic particles. The three important *sub* atomic particles you need to know about are protons electrons and *neutrons*. Protons and *neutrons* are found in the central **core** or nucleus of the atom. Electrons travel at high speed around the nucleus. Atoms are largely empty space except for the dense nucleus. As electrons orbit the nucleus, scientists thought that atoms might be like *mini* solar systems with the nucleus being like the sun and the electrons orbiting like *planets*. But electrons are not found in simple orbits like *planets*. There is an international colour **code** that is used when making or drawing space filling models of atoms or molecules.

This marked version of the text illustrates how words in the science-specific list occur in almost every line and are repeated in the text.

Comparing the GSL/AWL/science coverage from years nine to twelve

The coverage of the GSL/AWL/science/proper noun lists over the four individual textbooks is shown in Table Three below. Overall, we can see that the lists cover a higher percentage of the words at year nine (92.24% or 92.61% plus proper nouns), which decreases year by year to 87.82% at year twelve (88.5% with proper nouns). The year nine text contains more words from the GSL first and second thousand (78.85%). The coverage of these two lists decreases slightly each year, to year twelve at 76.89%. On average, the GSL lists cover 77%.

The AWL maintains a steady coverage over the four texts and averages 6.99%. The science-specific list is slightly lower than the AWL, with an average of 5.8%. This figure is around 2% higher than the coverage over an academic science corpus in Coxhead and Hirsh (2007). However, the average over years nine through eleven is 6.24%, whereas coverage over year twelve is 4.58%. Further, the year twelve text contains fewer family members of the AWL than the other textbooks do. It is important to note in Table Three above that the number of words not found in any list increases steadily from year nine (7.39 %) to year twelve (11.50). Note also that the proper noun coverage rises in this textbook by over half the coverage in years nine and ten, and almost double year eleven. These figures all suggest that the year twelve textbook has a different flavour to the other books.

Table Three: Coverage of the GSL/AWL/ Science-specific lists over the four individual textbooks

Text	Year 9	Running total	Year 10	Running total	Year 11	Running total	Year 12	Running total	Average
GSL 1000	71.03	71.03	69.88	69.88	69.64	69.64	70.08	70.08	70.1
GSL 2000	7.82	78.85	6.91	76.79	6.16	75.8	6.81	76.89	6.9
AWL	7.25	86.10	6.98	83.77	7.48	83.28	6.25	83.14	6.9
Science-specific list	6.14	92.24	6.45	90.22	6.13	89.41	4.58	87.82	5.8
Proper nouns	0.37	92.61	0.38	90.6	0.44	89.85	0.78	88.5	0.49
Not in any list	7.39	100	9.40	100	10.15	100	11.50	100	10.1

A final point to make about the data in Table Three is that just over 10% of the words in the textbooks do not occur in the four word lists used in this analysis. These words not found in any list include lexical items that are closely related to science. Thirteen of these words occur in all four textbooks over 100 times. These words are (most frequent first), *dioxide*, *crust*, *chromosomes*, *planet*, *genes*, *lens*, *sperm*, *offspring*, *gametes*, *gene*, *pole*, *friction*, and *sulphate*. Of these 13, only *genes*, *gametes*, and *pole* occur more than ten times in each textbook. The other words have more irregular patterns of occurrence.

To summarise, the four word lists plus proper nouns provide 90.39% coverage of the textbooks. This is reasonable coverage, but it means the existing word lists do not go far enough to cover the vocabulary that students need to read these books.

Research question two: How many words do learners need to know to read secondary science textbooks?

The purpose of this analysis is to find out the coverage of Nation's BNC lists over the textbooks and compare these data with earlier studies using the same lists. Table Four below shows the cumulative coverage of the BNC lists and the number of word families taken to reach that coverage. At the 14,000 list with proper nouns, the coverage of the textbooks by the BNC word families reaches 98.07%. The coverage of the proper noun list is just under 0.50%. It takes 4,274 word families to reach 98% coverage. The first 2,000 word families of the BNC lists reach 81.03% of the texts. If we compare this figure with coverage by the same lists over other corpora, we find it is close to the 83% over newspapers cited in Nation (2006, p. 72) but is lower than the 91% coverage of a graded reader in the same study (p. 73) and novels at approximately 88% (pp. 70-71).

Table Four: Coverage of all the science textbooks by the BNC lists

Word list	Tokens/ Percentage	Families
2,000	81.03	1602
4,000 + proper nouns	92	2851
9,000 + proper nouns	96.5	3842
14,000 + proper nouns	98.07	4274
Proper nouns	0.48	269
Not in the lists	1.23	????

How does the coverage of the BNC lists compare over years nine through twelve? Table Five below reports on 20,000 BNC lists and the proper nouns in each textbook. It shows that the coverage of the first 2,000 word families of the list is reasonably consistent across the textbooks, but drops two percent from year nine to year twelve. When we add proper nouns to the 4,000 list, we find coverage rises to 93.07% in year nine. Again, this coverage drops around two percent to 90.96% in year twelve.

Table Five: Text coverage of the four science textbooks by the BNC

Word list	Year 9	Year 10	Year 11	Year 12
2,000	82.27	81.29	80.53	80.09
4,000 + proper nouns	93.07	92.27	91.67	90.96
9,000 + proper nouns	97.28	96.81	96.32	95.52
11,000 + proper nouns	98.08	98.17	97.54	96.60
15,000 + proper nouns	98.72	98.74	98.17	97.18
Proper nouns	0.37	0.38	0.44	0.78
Not in the lists	0.87	0.75	1.33	2.08

We can see that 98% is reached at 11,000 plus proper nouns for years nine and ten, but at 15,000 plus proper nouns for the year 11 text. The year twelve textbook does not reach 98% because 2.08% of the words in the textbook do not occur in any of these word lists. Examples of words that do not occur in any list in the year twelve textbook include *sulphate*, *sulphur*, *gondwana* and *Gondwanaland*, as well as a number of the names of native New Zealand animals, including *tuatara* and *takahe*. Nation's (2006) study finds that 9,000 plus proper nouns are needed to reach 98% coverage of novels (p. 71), 8,000 plus proper nouns for newspapers (p. 72), and 3,000 plus proper nouns for a graded reader. Learners need a bigger vocabulary to read the science textbooks than they need to read novels and newspapers.

Here is the same sample of text we used to illustrate the occurrences of words in the GLS/AWL/science-specific lists above. This time, unmarked words in the text are in the first 1,000 words of the BNC, words marked with <2> are in the second thousand words, words marked with <3> are in the third thousand words, and so on. Words marked with <!> are not in any of the lists.

> <2>Model b <3>Mini <8>solar system
>
> As scientists <2>investigated <5>atoms more closely, they <2>discovered that <5>atoms were actually made of smaller <2>sub <5>atomic <5>particles. The three important <2>sub <5>atomic <5>particles you need to know about are <9>protons, <9>electrons and <10>neutrons. <9>Protons and <10>neutrons are found in the central <3>core or <6>nucleus of the <5>atom. <9>Electrons travel at high speed around the <6>nucleus. <5>Atoms are largely <2>empty space except for the <5>dense <6>nucleus. As <9>electrons <5>orbit the <6>nucleus, scientists thought that <5>atoms might be like <3>mini <8>solar systems with the <6>nucleus being like the sun and the <9>electrons <!>orbiting like <4>planets. But <9>electrons are not found in simple <5>orbits like <4>planets. There is an <2>international colour <2>code that is used when making or drawing space filling <2>models of <5>atoms or <6>molecules.

We can see that the 9,000 BNC list contains key words for this short text, such as *electrons* and *protons*.

These data show that to read these textbooks, a learner needs to know at least 3,000 more words than to read a novel in English. Furthermore, there is a considerable increase between 11,000 plus proper nouns in years nine and ten to over 20,000 plus proper nouns in year twelve.

Implications for teaching, materials design, and learning

How can teachers make use of the data from this study to help prepare their students for reading textbooks? Firstly, we can see that textbooks might vary considerably in terms of length and purpose, even if they come from the same series. The earlier textbooks have a different vocabulary than later textbooks in this series do. This means that students who might struggle with a year twelve text on a topic could benefit first from reading another version in an earlier textbook. The reader will encounter many of the same words in the texts but not so many words that are less common. Teachers can also use the results of an analysis with Range (BNC or GSL/AWL/science versions) to create two levels of text on the same topic. A less challenging text can be created using more frequent synonyms for difficult words, so readers can tackle an easier version first to explore ideas and language and then read a more challenging version.

Secondly, we can see the potential for word lists to help quickly identify the words in texts which might cause a challenge. The GSL/AWL/science lists plus proper nouns provide coverage of approximately 90% of the textbooks with 1798 fewer word families than the BNC lists. However, the BNC lists provide a higher coverage at 98% with 14,000 words plus proper nouns. Another benefit of the BNC lists is they illustrate how specialised science vocabulary can be spread across the different levels in the lists. Teachers could find out roughly how many words their students know using a vocabulary size test (see Nation, n.d.; Beglar, 2010) based on Nation's BNC lists. At the time of writing, the freely available test is up to 14,000 word families. Bilingual versions are also available on Nation's website in Korean, Mandarin and Vietnamese. Research is under way to develop and trial versions of the test up to 20,000 word families. The Vocabulary Levels Test (Schmitt, Schmitt & Clapham, 2001) contains the GSL first and second thousand and the AWL. (See Nation, 2001 or Coxhead, 2006 for versions of the test.)

Tools like Range can be used by teachers and students to investigate their own classroom texts and the results can be used to inform classroom material development. Both versions of Range are available on Paul Nation's website (see Heatley et al., 2003). Tom Cobb's Compleat Lexical Tutor website (n.d.) has a colourful interface using the Range programme, again with the two versions we used above. Another useful tool on the Cobb website is a concordance page. A teacher or student can input their own text which is then turned into an index of all the sentences in the text. By way of example, we found that the word *atomic* occurs 35 times in the year eleven textbook chapter on atoms. We also found that 25 of those occurrences were followed by the word *number* or *numbers*, as in the sentence, "Find the names of the elements whose atoms have the following atomic numbers". The other ten collocates of *atomic* are *atomic particle/s* and *atomic structure/s* (five each). Teachers can use these data to decide what other information is needed to understand the word *atomic* in light of its frequency and the words that co-occur.

The focus of this article is not teaching and learning vocabulary per se. We recommend developing activities to encourage vocabulary learning in and out of class. We also recommend students explore the content of the textbooks with a vocabulary focus. This means creating classroom activities that require learners to encounter and use words in a variety of ways, as Harmon et al. (2000) also suggested. See Nation (2008) for ways to develop activities such as split information, where learners have to share information to complete a task, and ranking activities, such as deciding which are the most important aspects of an atom, that focus on vocabulary. See Hirsh and Coxhead (2009) for ten suggestions of ways to focus on learning science-related vocabulary using Nation's (2007) four strands as a framework. These four strands are meaning-focused input, which involves learning from what students read and listen to, meaning-focused output, which involves writing and speaking, language-focused learning, which involves looking at different aspects of word knowledge such as how a word is spelled, pronounced, and used grammatically, and fluency. Coxhead (2006) contains ideas on activities for reading, writing, speaking and reading that focus on the word families of the AWL.

Limitations and further research

A benefit of studying current textbooks in use is that we can provide up-to-date findings for teachers and students. A drawback of such research is that textbooks might be rewritten and the results of a study might be out of date quite quickly. The publishers of the textbooks in this study, for example, have just published a new series of the same textbooks after our analysis. It will be interesting to see whether the vocabulary in those textbooks is the same as or different from the vocabulary in this present study.

Another difficulty in studying textbooks is selecting books which are used widely and regularly in classrooms. That is, ensuring that the textbooks are representative of the reading that students in secondary schools actually do in their studies. It might be the case that teachers and learners only use small sections of the textbooks or use materials from other sources in lessons. For example, teachers might incorporate online materials such as clips from YouTube into lessons. Electronic texts, such as e-books, make texts more readily available for corpus analysis and materials development. When possible, entire texts in electronic form should be used in research such as this instead of scanned texts, to avoid difficulties when scanning.

More analysis with longer texts would shed light on whether the short texts or the nature of the vocabulary in these texts caused this high coverage. Useful further research would include investigating ways to develop word lists that would be better ways to cover the vocabulary needed to read these texts. A much larger corpus of textbooks at secondary level is needed to find out more about the vocabulary of secondary textbooks. We also need to know how many words secondary school students actually know.

Conclusion

In this article, we have looked at the vocabulary in a series of four science textbooks from years nine through twelve using two different approaches. We have found that existing word lists are useful ways of finding out what vocabulary is being used in textbooks but they do not provide enough coverage. Nation's BNC lists show the spread of the vocabulary in these textbooks and show that EAL students need a reasonably large vocabulary to be able to read these books. We found a progression of lexical difficulty in the series from the books for junior and senior years. We have found that one of the reasons why secondary science textbooks might be difficult to read is that they possibly contain many unknown words.

Acknowledgement

The authors would like to acknowledge the support of the Victoria University Summer Research Scholarship, 2009/2010.

References

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

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101-118.

Biemiller, A. (2001). Teaching vocabulary: Early, direct, and sequential. *Essential readings on vocabulary instruction. The American Educator, 25*(1), 24-28

Bray, C., & Liew, J. (2008). *Survey of book publishing in New Zealand 2008.* Report Prepared for the Book Publishers Association of New Zealand. Wellington: Colmar Brunton.

Brown, C. (2007). Strategies for making social studies texts more comprehensible for English Language learners. *Social Studies, 98*(5), 185-188.

Chall, S. & Conrad, S. (1991). *Should textbooks challenge students*? New York: Teacher's College Press.

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

Cobb, T. (n.d.). 'The compleat lexical tutor', available at http://www.lextutor.ca/

Coxhead, A. (n.d.). The academic word list. Available at http://www.victoria.ac.nz/lals/resources/academicwordlist/default.aspx

Coxhead, A. (1998). *An academic word list.* English Language Institute Occasional Publication Number 18. Wellington: Victoria University of Wellington.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Coxhead, A. (2006). *Essentials of teaching academic vocabulary*. Boston: Heinle Cengage.

Coxhead, A. & Greene, J. (2009). Understanding and using an academic word list with middle and secondary level students. Regional Educational Laboratory Programme – Southeast Webinar, Thursday, August 27, 2009. Available at http://www.serve.org/Literacy/AWL.php.

Coxhead, A. & Hirsh, D. (2007). A pilot science word list for EAP. *Revue Française de Linguistique Appliqueé, XII*(2): 65-78.

Fang, Z. (2006). The language demands of science reading in middle school. *International Journal of Science Education, 28*(5), 491-520.

Greene, J. (2008). *Academic vocabulary and formulaic language in middle school content area textbooks*. Unpublished doctoral dissertation, Georgia State University, Atlanta.

Harmon, J., Hedrick, W., & Fox, E. (2000). A content analysis of vocabulary instruction in Social Studies textbooks for grades 4-8. *The Elementary School Journal, 100*(3), 253-271.

Heatley, A., Nation, P. & Coxhead, A. (2003). Range. Available at http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx.

Hirsh, D. & Coxhead, A. (2009). Ten ways of focusing on science-specific vocabulary in EAP. *English Australia Journal, 25*(1), 5-16.

Hook, G. (2004). *Science NCEA level one: Year 11*. Auckland: New House.

Hook, G. (2005). *Science Year 9*. Auckland: New House.

Hook, G. (2006). *Science Year 10*. Auckland: New House.

Hwang, K. (1989). *Reading newspapers for the improvement of vocabulary and reading*. Unpublished master's thesis, Victoria University of Wellington, New Zealand.

Martinez, I., Beck. S. & Panza, C. (2009). Academic vocabulary in agriculture research articles. *English for Specific Purposes, 28*, 183-198.

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

Nation, I.S.P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds). *Vocabulary in a second language: Selection, Acquisition and testing* (pp. 3-13). Amsterdam: John Benjamins.

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

Nation, I.S.P. (2007). The four strands. *Innovation in language learning and teaching, 1*(1), 1-12.

Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques.* Boston: Heinle Cengage.

Nation, I. S. P. and Webb, S. (2010). *Researching vocabulary.* Boston: Heinle Cengage.

Nurweni, A. and Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes, 18*, 172-193.

Relph, D., Croucher, M. & Castle, T. (2006). *Science Year 12: NCEA level two*. Auckland: Thomson New House.

Richards, J. (1974). Word lists: problems and prospects. *RELC Journal, 5*(2), 69-84.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88.

Taylor, C. (1979). *The English of high school textbooks.* Education Research and Development Committee report; No. 18. Canberra: Australian Government Publishing Service.

Webb, S. (2010). Using glossaries to increase the lexical coverage of television programs. *Reading in a foreign language, 22*(1), 201–221.

Webb, S. & Nation, P. (2009). Evaluating the vocabulary load of written text. *TESOLANZ Journal. 16,* 1-10.

Webb, S. & Rodgers, M. (2009). The vocabulary demands of television programs. *Language Learning*, *59*(2), 335-366.

West, M. (1953). *A general service list of English words.* London: Longman, Green and Co.