Reading in a Foreign Language April 2019, Volume 31, No. 1
ISSN 1539-0578 pp. 62–80

Can L1 children's literature be used in the English language classroom? High frequency words in writing for children

John Macalister
Victoria University of Wellington
New Zealand

Stuart Webb
Western University
Canada

Abstract

A challenge in reading research, and particularly extensive reading research, is how to manage the transition from the top of graded reading schemes to authentic texts which may be separated from each other by up to 5,000 word families. While texts written for native-speaker children have been recommended at times, recent research has shown that the lexical load of these texts was of similar difficulty to that of texts written for adults. In this paper we investigate whether it is possible to identify a specialist high frequency list in writing for children, and the impact of any such list on readability for language learners with a 2,000-word family vocabulary size. We found a list of 245 word families provided almost 3.4% coverage for such learners, thus making the use of L1 children's literature possible in the English language, and especially the English as a foreign language (EFL), classroom.

Keywords: extensive reading, graded readers, children's literature, vocabulary, imaginative writing, language learning, English as a foreign language

Developments in information technology over recent decades have spurred dramatic advances in the field of corpus linguistics. The one-million-word first-generation corpora that were significant achievements and contributors to our understanding of language in use at the time of their creation have now been dwarfed by vast monitor corpora. At the same time, as the creation of corpora has become easier, more and more specialized corpora have been created. The net result of these developments is that corpus linguists have been able to offer insights into many areas of language-related activity, with a particularly fruitful area being language learning and teaching (for a recent overview of the field, see Flowerdew, 2012).

One area of language learning and teaching where corpus linguistics has had a particular effect is that of vocabulary studies, through the production of word frequency lists. While it is true that such lists were in existence in pre-electronic corpora days (for a concise overview, see Kennedy,

1998, pp. 93–97) and some, such as the General Service List (or GSL; West, 1953), continue to be in use today, there have been considerable advances through electronic corpora. As one example, Nation (2006) created frequency lists for the 14,000 most frequent words using the British National Corpus, one of the new monitor corpora. As another example, with relevance to EAP teaching, Coxhead (2000) developed the Academic Word List, which claims to provide coverage of 10% of academic English through 570 word families. As even more specialized examples, there are the Nursing Education Word List (Mukundan & Jin, 2012), which also claims to provide coverage of almost 10% of the lexis in that field through learning a list of 969 technical words, and a 299-word engineering list for foundation level students (Ward, 2009). Other examples are a medical word list of 595 word families (Hsu, 2013) and an engineering word list of 729 word families which provided an impressive 14.3% coverage of the selected engineering textbooks (Hsu, 2014).

Literature Review

Lexical Coverage and Comprehension

The importance of such lists for language learning and teaching is that they provide information about lexical coverage (i.e., the percentage of known words in a text) and, indirectly, comprehension. There has been considerable interest in identifying the proportion of the lexis in a text that a person needs to know in order to read for comprehension (e.g., Hu & Nation, 2000; Laufer, 1989; Schmitt, Jiang, & Grabe, 2011), and the current agreement is that 98% of the vocabulary needs to be known for independent reading. Nation (2006) estimates this as an 8,000 to 9,000 word family vocabulary for comprehension of written text and a 6,000 to 7,000 word family vocabulary for spoken text (although recent research has found 95% coverage to be sufficient for comprehension of spoken text (van Zeeland & Schmitt, 2013), which would reduce the number of word families needed to be known). It is worth noting that these estimates are considerably higher than had earlier been thought. Previously 95% text coverage had been estimated as sufficient for independent reading comprehension (Laufer, 1989), and Hirsh and Nation (1992) estimated a vocabulary size of 5,000 word families was necessary for reading short novels. It should also be noted, first, that current estimates of vocabulary size may yet be revised (Schmitt, Cobb, Horst, & Schmitt, 2017) and, second, that judgements about vocabulary size and coverage may depend on what is regarded as adequate comprehension. For instance, Laufer and Ravenhorst-Kalovski (2010) suggest 95% coverage is sufficient for minimally acceptable comprehension (with 98% as optimal), and that "the 95% coverage can be achieved by 5,000 word families with proper nouns" (p. 26). This is a point that will be returned to in the Discussion section.

This does, of course, raise the question of what *comprehension* means. A useful way of considering this may be to differentiate between independent, instructional, and frustration levels of reading. Thus, if readers know 98% or more of the words on a page, they will be able to read independently and successfully understand. This should be the situation in extensive reading. Readers will still be able to read and successfully understand if they know 95% of the words on a page, but they will require instructional support (Grabe, 2009, p. 271). Instructional texts, Grabe (2009) goes on to say, should not go below the 90% level. Below that level readers reach the frustration stage. These levels all reinforce the importance of vocabulary knowledge to reading

comprehension. As Grabe (2009) has pointed out, "limited vocabulary knowledge may lead to a minimally developed and minimally coherent text model of comprehension" (p. 49). Furthermore, although compensatory reliance on a situation model may allow readers "to respond to a given task in a coherent way", this may not be in a way "that indicates comprehension of the text" (Grabe, 2009, p. 49). Or, to put it another way, top-down reading processes cannot compensate for limited bottom-up reading skills.

Approaches to Vocabulary Learning

If vocabulary is a key to comprehension – it has been described as the single most important predictor of success in reading (Laufer & Sim, 1985) – then a key question in language learning and teaching must be how to develop a vocabulary of sufficient size to allow successful, independent reading. A range of pedagogical approaches has been suggested, and these can usefully be divided into either direct or indirect approaches. Direct approaches involve deliberate teaching, and attention has been given to, among others, ways in which unknown vocabulary can be handled (Nation, 2004), the timing of attention given to new words (File & Adams, 2010; Sonbul & Schmitt, 2010), and activities that are learner- rather than teacher-directed, such as word cards (Webb, 2009a, 2009b). Pedagogical resources aimed at developing knowledge of specialized or technical vocabulary have also been developed and made commercially available (Howard, 2006, for example is the first of a series aimed at teaching the Academic Word List). Indirect, or incidental, learning tends to focus on reading, and extensive reading in particular. However, the vocabulary gains through extensive reading have been shown to be fragile (Waring & Takaki, 2003) and there has been debate as to the extent to which extensive reading alone can meet learners' vocabulary learning needs (Cobb, 2007; McQuillan & Krashen, 2008). The addition of tasks to extensive reading is one way in which vocabulary learning may be enhanced (Boutorwick, 2017; Green, 2005; Macalister, 2014). There has also been advocacy of narrow reading, following one topic over several texts, as a means of developing needed vocabulary (Gardner, 2008; Schmitt & Carter, 2000), which has also been translated into specific pedagogical activities (Watson, 2004).

The Importance of Repeated Exposure

In any vocabulary learning activity, repetition of the target item is essential. A single encounter with a new word is unlikely to lead to learning its form-meaning connection (Webb, 2007). Research investigating the effect of different levels of repetition has found that a minimum of ten encounters is needed for such learning to occur (Webb, 2007). In extensive reading repetition is ensured, both within and between texts, through graded readers, series of books written within a controlled vocabulary and also controlled for grammatical structure. For instance, in one well-known series Stage 1 begins with a vocabulary of 400 headwords, in texts of 5,200 words on average, with the past simple verb form singled out for attention; this contrasts with Stage 6, which employs a 2,500 headword vocabulary, in texts that average 30,000 words, and grammar that includes passives and advanced modal meanings.

The Transition from Graded to Authentic Reading

There is an issue, however, with developing a vocabulary of sufficient size (i.e., 8,000 or 9,000

word families for reading comprehension) that cannot easily be addressed by extensive reading of graded readers, which typically have an upper range in the vicinity of 3,000 headwords, or by direct teaching, given the constraints of time. The issue, in other words, is how to learn the many thousands of word families that remain unknown once a learner can read an upper level graded reader successfully and independently. In terms of vocabulary development through extensive reading, one suggestion has been that learners can read authentic children's literature, that is, texts written for young native speakers (Day & Bamford, 1998; Gardner, 2008; Mikulecky, 2009; Takase, 2009), and it is the case that such materials have been used successfully, including in the classic Fiji book flood (Elley & Mangubhai, 1981, 1983) and in languages other than English (Tabata-Sandom & Macalister, 2009). In a recent corpus-based study, however, Webb and Macalister (2013) concluded that the lexical load of texts written for native-speaker children were of similar difficulty to that of texts written for adults, and that neither was as well suited to extensive reading for language learners as graded readers. Webb and Macalister assumed a vocabulary of the 2,000 most frequent words in their study. In an earlier study that looked at materials for beginning readers and assumed a vocabulary of the 1,000 most frequent words, Jenkins (1993) also found that these authentic texts would be too difficult.

A more optimistic view of the transition from graded to authentic reading is offered by Uden, Schmitt, and Schmitt (2014). The optimism is partly based on an estimated smaller word family gap between the two (but still 3,000–4,000 word families; Uden et al., p. 18), and partly based on the results of a small-scale study in which three of four participants "made the jump to the ungraded novels without sacrificing much comprehension, reading speed, or satisfaction" (Uden et al., p. 19). However, these participants were highly motivated readers and their experience is not generalizable to less motivated readers, as the authors of the study themselves noted (Uden et al., p. 20).

Thus, given the advocacy of authentic children's literature, and the need to bridge the vocabulary gap from the upper limits of graded readers to authentic texts, it is worth considering whether a specialized vocabulary for this genre exists and, if it does, whether knowledge of those words would improve the readability of such texts for language learners. In his study, Jenkins (1993) found 216 frequently occurring word families beyond the 1,000 most frequent words of the GSL, and suggested that "as these word families are likely to occur in children's literature there will be considerable advantage gained by making sure they are known" (p. 108). Macalister (1999) reached a similar conclusion after an analysis of writing for more advanced young readers: "direct teaching of the [frequently repeated] word families above the 2,000 word level that are common both to Jenkins' list (1993: 134–142) [and this study] … would have an appreciable impact on readability and reading pleasure" (p. 73). It is worth noting, however, that both these studies were undertaken at a time when 95% text coverage and a 5,000-word family vocabulary were seen as sufficient for successful independent reading. It is also worth noting that there has not yet been a study to further Jenkins' (1993) and Macalister's (1999) claims for the desirability of a high frequency list for children's literature.

Research Questions

Considering the gaps in the literature outlined above, this study seeks to investigate the following three research questions:

1. What high frequency vocabulary beyond the 2,000 most frequent words of English can be identified from a corpus of children's literature?
2. If such high frequency vocabulary was known, what impact would it have on a second language learner's lexical coverage of authentic children's literature?
3. To what extent is such high frequency vocabulary representative of specialist language in children's literature?

Methodology

Two key considerations in any corpus-based study such as this are that the texts that form the corpus are representative of the phenomenon being investigated, and that the corpus is large enough to provide reliable information.

The Corpus

The texts chosen for this study come from the New Zealand *School Journal*, a resource for use in primary schools that has been published continuously since 1907. The *School Journal*, until changes in 2011, was published for many decades in four parts annually with each part targeting a different age group. Part One, for example, targets 7- to 8-year-olds, while Part Four is written for 11- to 13-year-olds. Each part appears in a number of issues each year, ranging from three to five. The *School Journals* contain informative and imaginative prose, as well as poetry and plays, and one way in which the readers are graded is by controlling text length. Thus, a Part One story (imaginative prose) is considerably shorter than a Part Four story. In creating the initial corpus, files tagged to identify the genre of the texts included were made for each Part. Also, during the corpus creation process, hyphens were removed from hyphenated words.

Text Selection

Given the focus of this study, only imaginative prose texts were included in the corpus. This decision was informed by the results of an earlier investigation of a small number of randomly selected imaginative and informative prose texts that suggested that imaginative prose passages would be suitable for extensive reading by second or foreign language learners at least in part because "the unknown words in imaginative prose are more likely to be repeated elsewhere within the corpus … than is the case for unknown words in informative prose passages" (Macalister, 1999, p. 80). This was supported by Gardner (2004, p. 24) who found narrative texts better for incidental vocabulary learning than expository. An informative prose or expository text may include numerous tokens of types specific to that particular content area (cf. findings about the effect of theme on vocabulary repetition in tightly themed expository non-fiction in Gardner, 2008), but non-specialized vocabulary common to story-telling is repeated in multiple imaginative prose texts. By focusing on multi-authored rather than sole authored texts, the author effect on vocabulary recycling is also removed (Gardner, 2008, although as he did not control for proper nouns in his study, at least some of the higher repetition in sole authored texts is accounted for by characters reappearing in different stories).

The focus on imaginative prose resulted in a corpus drawn from four years of publication of the

School Journal, comprising 174 texts totaling 128,540 tokens. This compares favorably with some other corpora used to investigate writing for children. Wharton (2005), for instance, had a corpus of just 1,871 tokens and even Baker and Freebody's corpus of 163 beginner readers used in the first two years of schooling was reasonably small, consisting of 83,838 tokens and 2,477 types (Baker & Freebody, 1989). Jenkins's (1993) corpus, also for beginning readers, was of a similar size, at 89,979 tokens. While there have been larger corpora of writing for children (Knowles & Malmkjær, 1996; Thompson & Sealey, 2007), the corpus created for this study did "reflect the size and shape of the documents from which it [was] drawn" (Sinclair, 1991, p. 19) and captured the complete publication of imaginative prose texts in the *School Journal* over a four year period.

Analysis

A preliminary analysis of the corpus was then made using RANGE (Nation & Heatley, 2002), which categorized words using the BNC 14,000 lists (Nation, 2006) and provided information about the occurrence of two other categories of words (proper nouns, and marginal words) as well as words not in any of the lists. Marginal words are such types as *aha* and *yuk*. As the assumption was that a reader would be familiar with the 2000 most frequent words in English, the results for the 3,000 to 14,000 lists were then examined to identify all word families with 10 or more repetitions, taking care to account for irregular verb forms, such as *sting* and *stung*. These high frequency word families beyond the 2,000 most frequent words thus became the focus for further analysis as the basis for a specialist word list for children's literature in order to address the first research question.

Word family was chosen as the unit of analysis because, in studies such as this where the focus is on receptive knowledge, the word family is "the most sensible unit" (Nation & Webb, 2011, p. 136). It assumes that a reader who knows one or two members of the word family should be able to recognize and understand other members of the family. The first six of the seven levels identified by Bauer and Nation (1993) are captured by the BNC word lists used in RANGE in this study (for further discussion of these issues, see Macalister & Webb, 2013).

In the process of analyzing the preliminary results, additional frequency and concordance data were obtained using Wordsmith to clarify questions such as word class (e.g., *bark* as a noun or a verb) and use (e.g., as a personal name). The Not in Lists list was also examined, and a small number of types were identified, such as *a-a-a-a-argh* and *arrrgggggghhhhhhh* as variant forms of the onomatopoeic marginal word *argh*, which were subsequently re-classified. Once these changes had been made, a modified RANGE including the specialist word list was again run, and the second research question addressed. Further information about changes made is given in the Results section of this paper.

Comparison and Validation

Two steps were taken to address the third research question. The first was a comparison with the high frequency vocabulary in writing for children identified by Jenkins (1993) who used a different corpus; the higher the degree of overlap, the more likely the new word list would be representative of the specialist vocabulary. The second comparison was to see the degree of

coverage offered by the new wordlist in corpora representing different genres. One corpus consisted of the complete *School Journal* publication for four years; another consisted of extracts from 33 graded readers; the third was texts drawn from the Wellington Written Corpus (Bauer, 1993) and were drawn from the Fiction and Press: Reportage sections.

Results

The initial examination of the corpus using RANGE revealed the distribution by frequency and other categories as shown in Table 1. The 2,000 most frequent words provided coverage of 87.35% of the corpus, and, as would be expected, the proportion of both tokens and types at each frequency level steadily decreased. The contribution of marginal words was smaller than might have been expected. Table 1 also shows the cumulative coverage. Thus, the 84.98% coverage in the first cell of that column includes the 1,000 highest frequency word families (81.29%) plus the proper nouns and marginal words (3.69%). It can be seen that 95% coverage (*) is achieved at the 4,000-word family level and 98% coverage (**) at the 8,000-word family level.

Table 1. *Coverage by frequency level and cumulative coverage, including proper nouns and marginal words, in the School Journal imaginative prose texts*

	Tokens	(%)	Cumulative coverage
Proper nouns	4528	(3.52)	-
Marginal words	221	(0.17)	-
1000	104,492	(81.29)	84.98
2000	7,785	(6.06)	91.04
3000	3,967	(3.09)	94.13
4000	1,762	(1.37)	95.50*
5000	1,404	(1.09)	96.59
6000	875	(0.68)	97.27
7000	583	(0.45)	97.72
8000	365	(0.28)	98.00**
9000	271	(0.21)	98.21
10000	183	(0.14)	95.35
11000	154	(0.12)	98.47
12000	87	(0.07)	98.54
13000	77	(0.06)	98.60
14000	69	(0.05)	98.65
Not in lists	1,716	(1.34)	99.99
TOTAL	128,539		

A list of types occurring 10 or more times in the 3,000 to 14,000 lists produced an initial list of 249 word families, all of which Wordsmith frequency ranking showed to be among the 1500 highest frequency words in the corpus. These 249 word families provided almost 4,400 tokens, or around 3.4% text coverage, and formed the basis for a new high frequency specialist wordlist. To do this, word families were removed from the assigned BNC list to a new baselist in RANGE. Further examination led to some adjustments in the above. Seven word families were removed as they were proper names (*Samoa*, *Zealand*) or used solely or primarily as personal names (*felicity*,

ginger, *harry*, *matt*, *minty*) as shown by Wordsmith concordances. Seventeen word families with 10 or more occurrences were identified in the Not in Any List file. In total these 17 word families accounted for over 23% of the Not in Any List tokens. These included proper nouns (e.g., *Cam*, with 27 tokens) and invented words (such as *freeble*, 62 tokens). In the end, only three of these word families (*boomerang*, *bubblegum*, *chirp*) were added to the initial word list of high frequency word families beyond the 2,000-word level. Once these changes had been completed, RANGE was run again with the results shown in Table 2.

Table 2. *Coverage by frequency level and cumulative coverage, including proper nouns and marginal words, in the School Journal imaginative prose texts following creation of specialist high frequency list*

	Tokens	(%)	Cumulative coverage
Proper nouns	4528	(3.52)	-
Marginal words	221	(0.17)	-
1000	104,492	(81.29)	84.98
2000	7,785	(6.06)	91.04
CH HF list	4,358	(3.39)	94.43
3000	1,773	(1.38)	95.81*
4000	1,027	(0.80)	96.61
5000	790	(0.61)	97.22
6000	584	(0.45)	97.67
7000	411	(0.32)	97.99
8000	249	(0.19)	98.18**
9000	191	(0.15)	98.33
10000	151	(0.12)	98.45
11000	133	(0.10)	98.55
12000	75	(0.06)	98.61
13000	66	(0.05)	98.66
14000	40	(0.03)	98.69
Not in lists	1,555	(1.21)	99.90
TOTAL	128,539		

The key difference between Tables 1 and 2 is the inclusion of the specialist wordlist as a separate line (the children's high frequency, or CH HF, list). The changes made between the first and second running of RANGE did not affect the coverage offered by the list. This remained at almost 4,400 tokens, or around 3.4% text coverage. The changes did, however, highlight the potential importance for reading children's literature that the list offers. Knowledge of the CH HF list means 95% coverage (*) is now achieved at the 3,000-word level. This point will be elaborated on in the Discussion section.

Some indication of the extent to which this list was representative of specialist high frequency vocabulary in writing for children was then gained by a comparison with Jenkins (1993); this is shown in Appendix A. Jenkins identified 66 word families beyond the 2,000-word level. Differences between the GSL in that earlier study and the BNC lists used in this study account for the fact that 24 word families that Jenkins identified as being beyond the 2,000-word level are included in the 2,000 high frequency word families of the BNC. However, a further 22 word

families were found to be high frequency in both studies, i.e., beyond the 2,000 word level of both lists; these formed part of the new list. An additional 13 word families that Jenkins identified were found to have multiple tokens (but fewer than 10) in the *School Journal* imaginative prose texts and only seven of Jenkins's word families were essentially unrepresented.

A further indication of the extent to which this list was representative of specialist high frequency vocabulary in writing for children can be gained by comparing its coverage of other genres. Table 3 shows the coverage offered by the CH HF list when the modified RANGE program was run on three corpora representing different genres and matched for size (285,143 tokens). One corpus consisted of the complete *School Journal* publication for four years; another consisted of extracts from 33 graded readers; the third was texts drawn from the Wellington Written Corpus (Bauer, 1993) and were drawn from the Fiction and Press: Reportage sections. As can be seen, in neither the graded reader nor the text written for adults' corpus does the new baselist achieve near the 3.39% coverage for imaginative prose shown in Table 3. It does achieve higher coverage in the *School Journal* corpus, which is partly explained by the fact that this corpus includes imaginative prose, as well as informative prose, poetry, and plays.

Table 3. *Coverage provided by specialist high frequency vocabulary in writing for children list*

Imaginative prose	*School Journal*	Graded Reader	Adult
3.39%	2.56%	0.8%	1.03%

Discussion

The first question that this study set out to investigate was whether it was possible to identify a list of specialist high frequency vocabulary in writing for children and, perhaps unsurprisingly, a corpus of *School Journal* imaginative prose texts did indeed yield such a list. The list is presented in Appendix B. Furthermore, the 245 word families can be classified into distinct categories. The 12 categories and the number of families in each are shown in Table 4. An additional 44 word families remained unclassified. Ten of the categories are semantic, and consist almost entirely of nouns. Some of the semantic categories that contributed to this list were intuitively likely, such as the language of school and storytelling and the inclusion of relatively closed categories such as the names of animals, clothing and parts of the body. The strong presence of verbs and, to a lesser extent, adjectives stood in strong contrast to the other ten, almost exclusively nominal, semantic categories.

Table 4. *Number of word families by category*

Category	*n*
Adjectives	17
Animal & Plant	26
Body	7
Clothing	9
Colour	2
Family	4
Food	12
House	15
Role	9
School	11
Storytelling	4
Verbs	85

The second question driving this study concerned the impact of the list on a second language learner's ability to read authentic children's literature. Offering almost 3.4% coverage, the 245 word families offer very substantial benefits in terms of bridging the gap between graded and authentic reading materials, a point that is returned to at the end of this section. This can be seen by considering that similar coverage (3.56%) would be gained by learning the 5,000 word families from the three to seven thousand word level (see Table 2).

The study also asked whether this high frequency vocabulary was unique to children's literature and, as shown in Table 3, the answer was in the affirmative, reinforced by the similarities found with Jenkins (1993). Thus, for language learners wanting to read beyond the upper level of graded readers, the CH HF wordlist offers a clear pathway to successful reading of children's literature. Indeed, the big difference in coverage between imaginative prose and graded readers (Table 3) shows that these word families will not be learned through reading graded readers, and so learning the word list is essential. Furthermore, learning words from the list may reflect typical L1 vocabulary development, given that they are words likely to be learned early by L1 speakers.

All the same, it is the case that 98% coverage remains at the 8,000-word level (Table 2), even if a mere 0.01% prevents 98% coverage at the 7,000-word level. In corpus-based studies such as this, however, it is necessary to go beyond what the raw numbers reveal, and to interrogate the data a little more closely as has been demonstrated in other studies (such as Baker, 2008; Harrington, 2008). Doing so suggests that an 8,000-word vocabulary size may not be necessary to achieve 98% coverage. One consideration, for instance, needs to be of the impact on coverage of the seven proper nouns mentioned earlier (*harry*, *ginger*, *felicity*, *minty*, *matt*, *Samoa*, *Zealand*). These are all found in the BNC lists beyond the 2,000 high frequency level, and each had 10 or more repetitions in the imaginative prose corpus, but remained in their original lists in the modified RANGE program. In other words, they did not form part of the new CH HF list. In

total, these seven word families contributed 144 tokens, equating to 0.11% coverage.

In a similar fashion, it is necessary to examine the contents of Not in Any Lists. As mentioned earlier, 17 word families with 10 or more occurrences accounted for over 23% of the tokens in that list. Further examination found numbers of compounds (e.g., *bluebird*, *sunhat*, *yessir*), variant spellings of both high (e.g., *b-b-but*, *ye-es*) and low frequency words (e.g., *stam-in-a*), proper nouns that had not been captured elsewhere (e.g., *Tongans* and 13 tokens of *Hoppy*, which form was not included in the *hop* word family that did enter the new baselist). It is also worth mentioning that Not in Any Lists included words of Maori origin (such as the bird *tui*, the sweet potato or *kumara*, and the word Maori itself) that would be familiar to a New Zealand-based readership (Macalister, 2006), as well as words of, for example, Greek or Samoan origin, often glossed. The presence of such words was, however, very small, and the key point here is that inclusion in Not in Any Lists does not necessarily equate with a high learning burden for a reader with a 2,000-word vocabulary.

Returning, then, to the need to bridge the vocabulary gap from the upper limits of graded readers to authentic texts, and the contribution that knowledge of a specialized vocabulary can make to improving the readability of such texts for language learners, the CH HF list can clearly make a significant contribution. With knowledge of the CH HF, however, a learner with a 2,000-word vocabulary is close to the 95%, and comfortably meets it if the learner is at the 3,000-level. At the 3,000-word level, then, a learner with knowledge of the CH HF is likely to be somewhere between Laufer and Ravenhorst-Kalovski's (2010) minimally acceptable and optimal comprehension levels. Remembering that the upper levels of graded reader schemes are typically around the 3,000-headword level, this suggests that authentic children's literature may be suitable reading material for such language learners.

Pedagogical Implications

The fact that a 245 word family list provides greater coverage than any lexical frequency band beyond the 2,000-word level, and that it can reduce the vocabulary size needed to achieve 95% coverage in writing for children by a one thousand word frequency band, suggests that the list deserves attention from language teachers. The amount of attention is likely to be affected by the language learning context, whether it is ESL or EFL. It seems intuitively likely that in an ESL setting learners would already be exposed to some of the word types contained in the high frequency list, such as language relating to school. It is in EFL contexts, therefore, that the CH HF list is likely to be most useful.

Despite the fact that some word types may already be familiar, learners would benefit from direct teaching of words on the list before beginning to read authentic writing for children. This may be particularly important for the large category of verbs and, to a lesser extent, that of adjectives. This is because a basic form-meaning link can usually be established pictorially with items in the other, almost exclusively nominal, semantic categories. Such form-meaning links are facilitated by the copious use of high-quality illustrations in the *School Journal*. These can be used to elucidate the text. As an example, the opening line of one story– "Greg's neighbour, Mr Forbes, was watching Greg on his skateboard"–is accompanied by a picture of a man watching a boy on

a skateboard (Berge, 1998). Identification of Greg, Mr Forbes, and the skateboard can quickly be established. 'Watching', on the other hand, requires greater explanation as the action cannot be captured in a concrete picture. However, it should not be necessary to devote a large amount of time to introducing the new words for, and despite Gardner's (2008, p. 109) suggestion to the contrary in relation to narrow fiction reading, the imaginative prose corpus does provide "increased repetitive exposure to [this list of] new or less familiar words."

Any direct teaching of the words in the list does, however, need to bear in mind the interference principle, which says that "the items in a language course should be sequenced so that items which are learned together have a positive effect on each other for learning, and so that interference effects are avoided" (Nation & Macalister, 2010, pp. 48–50). As the intention of the list is to reduce, not increase, the learning burden, this would mean in practice that formally related types such as *basket* and *bucket* should not be learned together, and nor should semantically related types such as *mouse* and *rat*, *mumble* and *mutter*, or *roar*, *scream*, and *yell*. On the other hand, types such as *spider* and *web* can support each other.

A further point to bear in mind is that this list, and the reading of imaginative prose, does not remove the necessity of learners acquiring appropriate academic vocabulary (cf. Gardner, 2004, p. 29). Such vocabulary is essential for developing subject-specific content knowledge and warrants explicit attention. The Academic Word List (Coxhead, 2000) remains the best available starting point for such attention; although based on written text, it has also been found to be "an effective tool to support listening to academic spoken English for different disciplines" (Dang & Webb, 2014, p. 73). Another option is the more recently-developed Academic Vocabulary List (Gardner & Davies, 2014).

Limitations and Future Directions

One question that should be asked of any corpus-based study such as this is whether the corpus is representative of the object of investigation, in this case writing for children. One possible limitation of the corpus used in this study may be that it is too context-specific, being drawn from writing for New Zealand children. As mentioned in the Discussion section, words that would be familiar to a New Zealand-based readership were present in the corpus but their presence was very small and did not appear to influence the words in the high frequency list (Appendix B).

Related to the question of representativeness, the fact that the high frequency list was drawn from informative prose texts in the *School Journal*, a sub-set of the *School Journal* corpus (Table 3), may have overstated the coverage offered. It was not possible to test the list coverage on another corpus of L1 children's literature. However, this remains a worthwhile undertaking for a later study.

A further limitation, a limitation shared by corpus-based studies in general, is that the impact of the high frequency list has not been trialled in the English language, and especially the EFL, classroom. This would also be a worthwhile focus for a later study. Such a study could establish baseline knowledge of the words in the list, and then investigate the impact of direct teaching of

the list on reading comprehension.

Conclusion

The aim of this paper has been to investigate whether it is possible to make the vocabulary load of reading authentic writing for children manageable for language learners with a 2,000-word vocabulary size. Examination of a corpus of imaginative prose for children has identified a relatively small high frequency list of 245 word families specific to this genre, some of which are likely to be familiar to learners through their inclusion in course books and from being encountered in the learners' immediate context. Depending on the way in which Not in Any List word families are regarded, this specialist list of high frequency vocabulary in writing for children has the potential to reduce the vocabulary size needed for successful reading of this genre by at least one 1,000-word family frequency band. It is, therefore, likely to make the vocabulary load of reading authentic writing for children manageable for language learners with a 2,000-word vocabulary size, even more so if they are able to read at the upper levels of graded readers successfully. Even more striking, 95% coverage is achieved at the 3,000-word family level, much lower than previous estimates. As a result, the CH HF list has the potential to assist learners' transition from the upper levels of graded readers to reading authentic texts, a transition that has challenged reading researchers for a considerable time, particularly in the extensive reading field. Given this potential, this specialist list deserves attention in the English language learning classroom.

References

Baker, C. D., & Freebody, P. (1989). *Children's first school books: Introductions to the culture of literacy*. Oxford, UK: Basil Blackwell.

Baker, P. (2008). 'Eligible' bachelors and 'frustrated' spinsters: Corpus linguistics, gender and language. In K. Harrington, L. Litosseliti, H. Saunston, & J. Sunderland (Eds.), *Gender and language research methodologies* (pp. 73–84). Basingstoke, UK: Palgrave Macmillan.

Bauer, L. (1993). *Manual of information to accompany the Wellington corpus of New Zealand English*. Wellington, New Zealand: Department of Linguistics, Victoria University of Wellington.

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

Berge, B. (1998). The sock gobbler. *New Zealand School Journal*, *1*(3), 2–9.

Boutorwick, T. J. (2017). *Vocabulary development through reading: A comparison of approaches* (Unpublished doctoral dissertation). Victoria University of Wellington, Wellington, New Zealand.

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, *11*(3), 38–64. Retrieved from http://llt.msu.edu/vol11num3/pdf/cobb.pdf

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

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*, 66–76. doi:http://dx.doi.org/10.1016/j.esp.2013.08.001

Day, R. R., & Bamford, J. (1998). *Extensive reading in the second language classroom*. Cambridge, UK: Cambridge University Press.

Elley, W. B., & Mangubhai, F. (1981). *The impact of a book flood in Fiji primary schools*. Wellington, New Zealand: New Zealand Council for Educational Research.

Elley, W. B., & Mangubhai, F. (1983). The impact of reading on second language learning. *Reading Research Quarterly*, *19*, 53–67.

File, K. A., & Adams, R. (2010). Should vocabulary instruction be integrated or isolated? . *TESOL Quarterly*, *44*(2), 222–249. doi: 10.5054/tq.2010.219943

Flowerdew, L. (2012). *Corpora and language education*. Basingstoke, Hampshire: Palgrave Macmillan.

Gardner, D. (2004). Vocabulary input through extensive reading: A comparison of words found in children's narratives and expository reading materials. *Applied Linguistics, 25*, 1–37. http://dx.doi.org/10.1093/applin/25.1.1

Gardner, D. (2008). Vocabulary recycling in children's authentic reading materials: A corpus-based investigation of narrow reading. *Reading in a Foreign Language, 20*, 92–122.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*, 305–327. doi:10.1093/applin/amt015

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge, UK: Cambridge University Press.

Green, C. (2005). Integrating extensive reading in the task-based curriculum. *ELT Journal, 59*, 306–311. https://doi.org/10.1093/elt/cci059

Harrington, K. (2008). Perpetuating difference? Corpus linguistics and the gendering of Reported dialogue. In K. Harrington, L. Litosseliti, H. Saunston, & J. Sunderland (Eds.), *Gender and language research methodologies* (pp. 85–102). Basingstoke: Palgrave Macmillan.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8*, 689–696.

Howard, J. (2006). *College vocabulary 1*. Boston: Houghton Mifflin.

Hsu, W. (2013). Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research*, *17*, 454–484. doi:10.1177/1362168813494121

Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes, 33*, 54–65. doi: 10.1016/j.esp.2013.07.001

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*, 403–430.

Jenkins, S. (1993). *The vocabulary burden of controlled and uncontrolled reading materials used with beginning ESL readers* (Unpublished master's thesis). Victoria University of Wellington, Wellington, New Zealand.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London, UK: Longman.

Knowles, M., & Malmkjær, K. (1996). *Language and control in children's literature*. London and New York: Routledge.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Laurén & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Clevedon, UK: Multilingual Matters.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 2*, 15–30.

Laufer, B., & Sim, D. D. (1985). Measuring and explaining the reading threshold needed for English for academic purposes texts. *Foreign Language Annals, 18,* 405–411. doi:10.1111/j.1944-9720.1985.tb00973.x

Macalister, J. (1999). The School Journals and TESOL: An evaluation of the reading difficulty of School Journals for second and foreign language learners. *New Zealand Studies in Applied Linguistics, 5,* 61–85.

Macalister, J. (2006). The Maori presence in the New Zealand English lexicon, 1850–2000: Evidence from a corpus-based study. *English World-Wide, 27,* 1–24.

Macalister, J. (2014). The say-it activity. *Modern English Teacher, 23*(1), 29–32.

Macalister, J., & Webb, S. (2013). A response. *TESOL Quarterly, 47,* 852–855. doi:10.1002/tesq.142

McQuillan, J., & Krashen, S. D. (2008). Commentary: Can free reading take you all the way? A response to Cobb (2007). *Language Learning & Technology, 12*(1), 104–108.

Mikulecky, L. J. (2009). Using Internet-based children's and young adult literature for extensive reading in EFL instruction. In A. Cirocki (Ed.), *Extensive reading in English language teaching* (pp. 333–347). Munich, Germany: Lincom.

Mukundan, J., & Jin, N. Y. (2012). Development of a technical nursing education word list (NEWL). In H. P. Widodo & G. G. Park (Eds.). *Moving TESOL beyond the comfort zone: Exploring criticality in TESOL* (pp. 81–100). Hauppauge, NY: Nova Science Publishers.

Nation, I. S. P. (2004). Vocabulary learning and intensive reading. *EA Journal, 21*(2), 20–29.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63,* 59–82. doi: 10.1353/cml.2006.0049

Nation, I. S. P., & Heatley, A. (2002). RANGE: A program for the analysis of vocabuary in texts. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Nation, I. S. P., & Macalister, J. (2010). *Language curriculum design*. New York and London: Routledge/Taylor & Francis.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle, Cengage Learning.

Schmitt, N., & Carter, R. (2000). The lexical advantages of narrow reading for second language learners. *TESOL Journal, 9,* 4–9. doi:10.1002/j.1949-3533.2000.tb00220.x

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching, 50,* 212–226. doi:10.1017/S0261444815000075

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95,* 26–43. doi:10.1111/j.1540-4781.2011.01146.x

Sinclair, J. (1991). *Corpus concordance collocation*. Oxford, UK: Oxford University Press.

Sonbul, S., & Schmitt, N. (2010). Direct teaching of vocabulary after reading: is it worth the effort? *ELT Journal, 64,* 253–260. doi:10.1093/elt/ccp059

Tabata-Sandom, M., & Macalister, J. (2009). That 'eureka feeling': A case study of extensive reading in Japanese. *New Zealand Studies in Applied Linguistics, 15*(2), 41–60.

Takase, A. (2009). The effects of different types of extensive reading materials on reading amount, attitude and motivation. In A. Cirocki (Ed.), *Extensive reading in English language teaching* (pp. 451–465). Munich, Germany: Lincom.

Thompson, P., & Sealey, A. (2007). Through children's eyes? Corpus evidence of the features of children's literature. *International Journal of Corpus Linguistics, 12,* 1–23.

Uden, J., Schmitt, D., & Schmitt, N. (2014). Jumping from the highest graded readers to ungraded novels: Four case studies. *Reading in a Foreign Language*, *26*, 1–28.

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*, 457–479. doi:10.1093/applin/ams074

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, *28*, 170–182. https://doi.org/10.1016/j.esp.2009.04.001.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*, 130–163.

Watson, J. (2004). Issue logs. In J. Bamford & R. R. Day (Eds.), *Extensive reading activities for teaching language* (pp. 37–39). Cambridge, UK: Cambridge University Press.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*, 46–65. doi:10.1093/applin/aml048

Webb, S. (2009a). The effects of pre-learning vocabulary on reading comprehension and writing. *Canadian Modern Language Review*, *65*, 441–470. doi: 10.3138/cmlr.65.3.441b

Webb, S. (2009b). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal*, *40*, 360–376. https://doi.org/10.1177/0033688209343854

Webb, S., & Macalister, J. (2013). Is text written for children useful for L2 extensive reading? *TESOL Quarterly*, *47*, 300–322. https://doi.org/10.1002/tesq.70

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

Wharton, S. (2005). Invisible females, incapable males: Gender construction in a children's reading scheme. *Language and Education*, *19*, 238–251.

Appendix A

Word Families beyond the GSL 2,000 High Frequency List Identified by Jenkins (1993)

BNC 2000 words that are present in the *School Journal* lists but were not among the 2000 high frequency words in the GSL (*n* = 24)	Words beyond the BNC 2000 common to both, and added to the new baselist (*n* = 22)	Words identified by Jenkins with fewer than 10 tokens in *School Journal* lists (*n* = 13)	Words identified by Jenkins but with no or single tokens in *School Journal* lists (*n* = 7)
bang	ant	bark	crocodile
birthday	banana	canoe	fairy
biscuit	beach	claw	fuss
chase	blanket	delicious	hare
chip	bubble	fiddle	mosquito
chocolate	cheek	flap	pumpkin
chop	clap	fox	scrap
enormous	frog	sausage	
foot	ghost	terror	
horrible	gobble	trot	
icecream	hedge	tug	
jacket	honey	witch	
kid	hop	wolf	
mum	hug		
naughty	lamb		
ok	lick		
plate	lion		
pop (v)	magic		
sack	rooster		
scared	snap		
silly	stare		
tiny	yell		
trousers			
vegetable			

Appendix B

Writing for Children High Frequency (CH HF) Wordlist

Adjectives
angry
awesome
crazy
faint
fierce
fluffy (word family: fluff)
gentle
lean
neat
nervous
pale
silent (word family: silence)
smooth
sore
sticky
thirst
wild

Animals & Plants
ant
bull
cage
crab
creature
dragon
flea
frog
goat
holly
insect
jasmine
kitten
lamb
leaf
lion
mouse
paw
pet
pine
pup
rat
roost
seed
spider
web

Body
ache
blonde
breath
cheek
stomach
throat
wrist

Clothing
sleeve
greatcoat
helmet
jersey
jumper
shorts
sweatshirt
togs
towel

Colours
silver
purple

Family
cousin
grandad
nanny
papa

Food
banana
bubblegum
coconut
cookie
honey
jelly
lemonade
lolly
mushroom
noodle
spice
watercress

House
basket
blanket
broom
bucket
cushion
dishwasher
doorway
jar
ladder
lawn
lid
matchbox
oven
pillow
saucer

Roles
burglar (word family: burgle)
captain
emperor
pharaoh
pilot
princess
rabbi
soldier
vet

School
bat
bench
cardboard
cricket
gang
glue
lunchtime
notebook
playground
skateboard
soccer

Story
alien
ghost
giant
magic

Verbs
bounce
burst
carve
chew
chirp
clap
crash
crawl
creep
crouch
curl
dart
dive
drift
drip
flash
flick
fold
frown
gasp
giggle
glance
glare
glitter
glow
gobble
grin
groan
gulp
hiccup
hiss
hop
hug
hum
illustrate
kiss
leap
lick
moan
mow
mumble
mutter
nod
paddle
pause
peer
poke
protest
puff
rip
roar
scatter
scoop
scramble
scratch
scream
shine
shiver
shove
shrug
sigh
snap
sneak
sniff
spin
spray
stare
steal
strap
stroke
suck
surf
swallow
sweat
sweep
swing
thump
tuck
wag
wail
wander
whisper
wriggle
yell
zoom

Other	bush	junk	puddle	taxi
balloon	concert	lake	reward	tent
bandage	ditch	liquid	rope	tide
beach	gum	ms	rune	torch
bead	hammer	mud	shadow	trail
boomerang	heap	olympic	shelter	trailer
brand	hedge	paddock	spaceship	trap
bubble	hippy	pedal	storm	wart
bunch	hut	planet	string	yuk

About the Authors

John Macalister is Professor of Applied Linguistics at Victoria University of Wellington, New Zealand. His research & teaching interests include language learning and teaching, language teacher education, and language curriculum design. Email: john.macalister@vuw.ac.nz

Stuart Webb is Professor of Education at Western University, Canada. He conducts research on second language acquisition, particularly incidental vocabulary learning through reading, listening, and watching television, as well as how words can be taught effectively. Email: swebb27@uwo.ca