A reassessment of frequency and vocabulary size in L2 vocabulary teaching[1]

Norbert Schmitt

University of Nottingham, Nottingham, UK
norbert.schmitt@nottingham.ac.uk

Diane Schmitt

Nottingham Trent University, Nottingham, UK
diane.schmitt@ntu.ac.uk

Abstract

High-frequency vocabulary has traditionally been thought to consist of the 2,000 most frequent word families in English, and low-frequency vocabulary as that beyond the 10,000 frequency level. This paper argues that these boundaries should be reassessed on pedagogic grounds. Based on a number of perspectives (including frequency and acquisition studies, the amount of vocabulary necessary for English usage, the range of graded readers, and dictionary defining vocabulary), we argue that high-frequency vocabulary should include the most frequent 3,000 word families in English. We also propose that low-frequency vocabulary boundary should be lowered to the 9,000 level, on the basis that 8,000-9,000 word families is sufficient to provide the lexical resources necessary to be able to read a wide range of authentic texts (Nation 2006). We label the vocabulary between high-frequency (3,000) and low-frequency (9,000+) as MID-FREQUENCY vocabulary. We illustrate the necessity of mid-frequency vocabulary for

proficient language use, and make some initial suggestions for research addressing the pedagogical challenge raised by mid-frequency vocabulary.

Norbert Schmitt is Professor of Applied Linguistics at the University of Nottingham and is interested in all aspects of second language vocabulary studies. His current interests include vocabulary testing and the relationship between reading and vocabulary. His most recent book is *Researching Vocabulary*, a vocabulary research manual published with Palgrave Macmillan (2010).

Diane Schmitt is a Senior Lecturer in EFL/TESOL at Nottingham Trent University. She teaches on the MA in English Language Teaching and also on a range of English for Academic Purposes courses. She has co-authored textbooks on teaching academic vocabulary. She regularly presents, publishes and consults on the following areas: academic writing, plagiarism, vocabulary acquisition, language testing, materials development and the international student experience.

A Reassessment of Frequency and Vocabulary Size in L2 Vocabulary Teaching

1. Introduction

Frequency has long informed the principled selection of vocabulary[2] in L2 teaching pedagogy. Paul Nation, a long-term exponent of this approach, breaks vocabulary into four categories: high-frequency words, academic words, technical words, and low-frequency words (e.g. 2001a: 11-12; 2011: 12-13). His basic message, most recently reiterated in a 2011 'Thinking Allowed' piece, is that teachers and materials writers essentially need to make a cost/benefit analysis of vocabulary to decide whether or not any particular lexical item merits instruction/inclusion (see also Nation, 2001b). High-frequency vocabulary is extremely useful for learners, and so should be explicitly addressed. Academic vocabulary is worth focusing on for learners wishing to study in English, and the same goes for technical vocabulary for learners focusing on specific purpose domains. Conversely, in Nation's view, low-frequency vocabulary occurs so infrequently that it is not worth spending classroom time on these words. Rather, teachers should teach vocabulary learning strategies to learners, so they can learn these rarer words on their own.

While we agree with the cost/benefit approach, we feel that recent research has made the four-part categorization untenable as a pedagogic description. The key evidence is a more recent study by Nation (2006), in which he uses a solely frequency-based approach instead of the four-part categorization. In it, he calculates that it takes knowledge of 8,000-9,000 word families to read a diverse range of authentic texts in

English without unknown vocabulary being a substantial handicap. This vocabulary size takes us far beyond high-frequency vocabulary; in fact it takes us beyond current definitions of high-frequency, academic, and technical vocabulary combined. If it takes this much vocabulary for proficient English use, then clearly there needs to be a focus on vocabulary beyond that covered by the high-frequency, academic, and technical categories.

If frequency-based descriptions of English are to be of value to language practitioners, the extent and boundaries of high- and low-frequency vocabulary need to be carefully defined. High-frequency vocabulary has traditionally been operationalized as around the first 2,000 most frequent word families[3] in English. Conversely, low-frequency vocabulary has been characterized in various ways: ranging from anything beyond 2,000 word families all the way up to all of the word families beyond the 10,000 frequency level. However, it is unclear whether these traditional boundaries (which were never established in a rigorous manner) are set at the optimal levels, especially given Nation's higher vocabulary size targets (i.e. 8,000-9,000 word families for independent proficient use).

Frequency-based descriptions of English will also have to consider how to conceptualize the many thousands of word families which come between the high-frequency level and Nation's 8,000-9,000 family target (i.e. do the academic and technical categories cover these thousands of families?). One of the problems is that in discussions of frequency, general vocabulary is usually discussed in terms of 1,000 word categories of decreasing frequency. However, academic and technical vocabulary are subsets of general English which cut across these 1,000 word bands, and the four-part

categorization stemming from Nation's early work does not take account of this. Thus, when analyzing texts or planning what to teach, it is important to recognize that the notions of academic/technical vocabulary do not necessarily fill the gap between high- and low-frequency bands.

This paper attempts to address these issues by revisiting the scope of both high- and low-frequency vocabulary from multiple perspectives and suggesting new boundaries for each which make better sense in terms of what learners can do with various vocabulary sizes. It will then explore the vocabulary between the high- and low-frequency levels, and argue that the academic and technical categories do not adequately cover it. We introduce a new 'MID-FREQUENCY' category to describe this in-between frequency band, illustrate the benefits of knowing words in this band, and argue that these words need to be addressed in a principled way in language pedagogy.

2. What is high-frequency vocabulary?

The first 2,000 word families is the traditional cut-point for high frequency vocabulary, and is widely cited in teacher guidebooks and research publications (e.g. Nation 1990, 2001a; Read, 2000; Schmitt 2000; Thornbury, 2002). In this section, we will look at the origins of this figure and explore whether it is still appropriate. The 2,000 figure largely comes from the influence of the *General Service List (GSL)* (West 1953). The GSL includes a little over 2,000 headwords (essentially word families) and has been an important resource for teachers and material writers for many decades. The 2,000 figure was reinforced by research on oral discourse. Schonell, Meddleton, & Shaw (1956)

studied the speech of Australian workers, and found that approximately 2,000 word families covered around 99% of this discourse. It was thus concluded that around 2,000 word families were sufficient to engage in daily conversation. Based on this historical background, Nation set the initial frequency level for both his influential vocabulary research tool (*VocabProfiler*, see the 'Classic VP' on the *Lextutor* website <http://www.lextutor.ca/>) and his widely-used vocabulary test (*Vocabulary Levels Test,* Nation 1990) at 2,000 families, further reinforcing this level as the established initial stage of vocabulary, and by default, high-frequency vocabulary.

As can be seen, the origins of the 2,000 figure largely come from frequency counts and research which is over 50 years old. Given the increase in vocabulary research over the past 20 years, it seems reasonable to revisit the frequency issue to determine whether 2,000 is still the best boundary for high-frequency vocabulary, or whether an adjusted figure would prove more useful. We will explore this issue from a number of perspectives including frequency, coverage, acquisition, and use.

2.1 Frequency evidence

The first type of evidence to explore is the nature of the frequency distribution of vocabulary. It is well-known that a small number of word types occur very frequently and make up the majority of running words in discourse. Conversely, a very large number of types occur very rarely, and make up the remainder of running words. This is illustrated in Table 1 and Figure 1 which looks at Nation's (2006) analyses of nine written and spoken corpora (e.g. Brown, Kohlapur, Wellington written, and LUND corpora); the

general shape of these distributions would be similar for most other corpora. The written corpora include texts from sources such as novels and newspapers, while the spoken corpora include speech from sources such as everyday conversation with friends and family and people calling into radio programs.

(Table 1 and Figure 1)

For our discussion, the key feature of Table 1 and Figure 1 is the rapidly declining coverage obtained as vocabulary becomes less frequent. The first 1,000 word families clearly do the bulk of the work in English (in large part due to the extremely high frequency and coverage of function words). The second 1,000 contributes a much smaller, but still useful, amount of coverage, as does the third 1,000 to a lesser extent. But by the fourth 1,000 families, the coverage drops substantially, with only a maximum of 3% for 2,000 families (4^{th} and 5^{th} 1,000). Beyond this, the coverage return gets increasingly small. It could be argued that high-frequency vocabulary is that which occurs before the coverage percentages become so small that it is unlikely that the words will occur frequently across a wide range of texts. There is not a clearly identifiable cut-point (unless we limited high-frequency vocabulary to the first 1,000), but looking across a range of corpora (See Tables 1-3) frequency distributions show that beyond the 2,000-3,000 frequency levels, frequency of occurrence drops off to low levels. This suggests that high-frequency vocabulary would include the most frequent 2,000-3,000 word families in English.

2.2 Frequency and incidental acquisition

Further insight is provided by a small frequency/acquisition study carried out by Cobb (2007). He was interested in whether vocabulary at various frequency levels occurred often enough that it could be learned merely from incidental exposure (on the generous operating assumption that six occurrences were sufficient). He looked at 30 target words (10 from each of the 1,000, 2,000, and 3,000 levels) to see how often they occurred in a 517,000-word extract of the Brown written English corpus (divided into three types of discourse: press, academic, and fiction). He found that at least eight out of the ten target words from the first 1,000 and seven from the second 1,000 frequency levels occurred 6+ times. At the third 1,000 level, this dropped to between 3 (academic) – 5 (fiction) words. This suggests that the 3,000 level is the lowest frequency which we can consider 'high-frequency' in terms of learning opportunities from reading, and even then it is starting to become marginal. Cobb also assembled a 300,000-word corpus of novels from the author Jack London, and found that only 469 (57%) of the 817 3,000 level word families occurred six times or more, further illustrating that at the 3,000 level, learning opportunities begin to taper off quickly. This situation would deteriorate even further for word families at the 4000 and 5000 levels and beyond.

2.3 Frequency and use

We can also look at the frequency issue from the very practical standpoint of the amount of vocabulary necessary to function in English. In terms of high-frequency vocabulary,

this relates to the ability to use English at the basic, but still useful, end of the proficiency continuum. (We will address higher levels of proficiency in our discussions of low- and mid-frequency vocabulary below). Little work has been done on the lexical requirements for the productive skills (speaking, writing), but a small number of studies have been carried out on reading and listening. If learners wish to read a wide range of authentic novels or newspapers without assistance, then Nation (2006) calculates that it takes knowledge of the most frequent 8,000-9,000 word families to cover 98%[6] of this type of text, based on his wordlists derived from the British National Corpus (BNC). (Note that this does not mean a total vocabulary size of 8,000-9,000 word families, but rather good knowledge of the word families up to these specific frequency bands. A learner's total vocabulary size may include some word families beyond these frequency bands.) If we allow for lower comprehension expectations and use a less stringent coverage figure of 95%, this would still entail knowledge of word families up to the 4,000-5,000 frequency bands plus proper nouns (Laufer & Ravenhorst-Kalovski 2010). Even this lower figure would appear well beyond any reasonable definition of high-frequency vocabulary, and so it seems that reading a range of authentic texts is not possible with high-frequency vocabulary alone. However, reading would still be possible using graded readers (see below).

Listening at a conversational level (e.g. listening to narrative stories) appears to require a lexical coverage of only 95%[7] (van Zeeland & Schmitt under review), and this entails a vocabulary size of between 2,000-3,000 families. For example, Adolphs & Schmitt (2003) found that it took a little over 2,000 word families to reach 95% coverage of the five-million-word CANCODE[8] corpus, and around 3,000 individual word forms to

reach 95% coverage of the 4.2-million-word conversational sub-section of the spoken component of the BNC. Nation's (2006) analysis of approximately 200,000 words of unscripted speech in the Wellington Corpus of Spoken English showed that 3,000 word families plus proper nouns achieved a coverage of 96+%. Webb and Rodgers (2009a) analyzed the language of 88 television programs and found that knowledge of the most frequent 3,000 word families (plus proper nouns and marginal words (*oh, uh, mmm,* and *ah*)) provided 95.45% coverage. (This ranged from 2,000 to 4,000 word families in different TV genres). They also analyzed 318 film scripts (2009b) and found that the most frequent 3,000 word families provided 95.76% coverage (the range was 3,000 to 4,000 word families depending on the movie genre). Taken together, it seems that knowledge of the most frequent 3,000 word families should provide the lexical resources to largely understand (and presumably produce) conversational English. This vocabulary size may still be too small to enable full comprehension and enjoyment, but it seems adequate to make listening texts accessible enough to be useful for many purposes, including using texts for learning English. Overall, if aural competency is believed to be a basic language skill, then this evidence supports the argument for considering the first 3,000 word families as high-frequency vocabulary.

2.4 Graded readers

While we have seen that reading authentic texts requires a wider vocabulary than just high-frequency vocabulary, graded readers offer a pathway to begin reading with more limited lexical resources. There are a number of graded reader series offered by various

publishers, generally beginning at the 200 – 400 word level, and topping out at around 3,000-3,800 words. (The last stage level in the Oxford series gets up to the 5,000 level.) For example:

- *Macmillan Readers*: 300 – 2,200 headwords
- *Heinle Cengage Page Turners:* 200 – 2,600 headwords
- *Penguin Readers:* 200 – 3,000 headwords
- *Cambridge English Readers*: 400 – 3,800 headwords
- *Oxford Progressive English Readers:* <1,400 – 5,000 headwords

The fact that most graded reader series finish at around the 3,000 word family level implies that a vocabulary size of 3,000 word families is an important stage for ESL learners. However, as Tom Cobb (personal communication) notes, graded reader schemes seldom rely in any disciplined way on word frequency for their levels, but rather rely on the much looser idea of total number of headwords. For example, Oxford-Bookworms' *Elephant Man* is described as containing 400 headwords, but Cobb's informal *Lextutor* analysis shows that only about three-quarters of headwords (families) come from the first 1,000 frequency band, with the rest being widely distributed through the 2,000-9,000 frequency bands. Still, despite the lack of a consistent frequency procedure among graded readers, the point remains that in terms of vocabulary size, 3,000 families seems to be a key figure. As such, it remains a reflection of the basic vocabulary of English, and by extension, informs what might usefully be considered high-frequency vocabulary.

2.5 Lexicography and dictionary defining vocabulary

Dictionaries are a key lexical resource, giving access to a vast number of lexical items, but the monolingual dictionaries produced for native speakers can be difficult for learners to use, simply because the vocabulary in the definitions can often be as difficult as the word being looked up. Lexicographers producing learner dictionaries have considered this problem, and a typical solution is to create lists of DEFINING VOCABULARY, with which all of the entries in the dictionary are defined. The words selected for inclusion in these defining lists are judged to have particular utility for describing a wide variety of meanings, and are typically the highest frequency vocabulary in English. The extent of these defining vocabulary lists can give some indication of both 1) the most important vocabulary in English, and 2) the extent of the vocabulary which learners need to know towards the beginning of their studies in order to effectively use English-medium learner materials. The lists range from about 2,000 – 3,000 words depending on the publisher, for example:

- *Longman Dictionary of Contemporary English* (2009) ≈ 2,000
- *Oxford Advanced Learner's Dictionary* (2010) = 3,000
- *Macmillan English Dictionary for Advanced Learners* (2002) ≈ 2,500

A *Lextutor* analysis of these defining vocabulary lists shows that over 90% of their contents come from the first 3,000 most frequent word families, and over 95% from the first 4,000 families. This confirms that word utility (as judged by a variety of lexicographers) is very strongly related to high word frequency. If we accept that the most useful and widely-applicable vocabulary is largely captured by these defining

vocabularies (which correspond strongly with frequency), this suggests that the first 2,000 – 3,000 word families provide a workable definition of high-frequency vocabulary.

2.6 Defining high-frequency vocabulary

The goal of this section was to determine the most useful parameters of high-frequency vocabulary. The traditional boundary of high-frequency has been 2,000 word families, but according to most of the above perspectives, this seems too low. On balance, it seems that 3,000 word families is a more pedagogically-useful criterion. While learners can obviously communicate to some extent with much smaller vocabulary sizes than this, it appears that 3,000 word families represent an important milestone in language development. More vocabulary than 3,000 word families would allow learners to communicate in a wider range of situations, but the rapid decay in frequency of occurrence (Table 1 and Figure 1) makes it very difficult to consider vocabulary beyond the 3,000 level as 'high-frequency'. Therefore, we propose that the first 3,000 word families of English be considered high-frequency (and thus maximally-useful) vocabulary. As Cobb (2007: 41) observes, 'The first three of Nation's BNC lists (i.e. the 3000 most frequent word families) represent the current best estimate of the basic learner lexicon of English.'. The evidence presented here provides a sound basis for setting the upper limit of high frequency vocabulary at the 3,000 most frequent word families.

3. What is low-frequency vocabulary?

We now look at the other extreme of the frequency continuum, where vocabulary becomes so infrequent that it has very limited utility. The obvious way of setting the boundary of low-frequency vocabulary is by looking at frequency distributions. However, while the nature of the frequency distribution of English words makes it feasible to suggest a reasonable cut-point for high-frequency vocabulary, this is not the case for low-frequency vocabulary. Nation (2006) used the first fourteen 1,000 level frequency bands from the BNC to determine the percentage of coverage across nine spoken and written corpora (See Table 1). Table 2 illustrates his results for just the Lancaster-Oslo-Bergen (LOB) 1-million-token corpus of written British English, but the other corpora produced similar results. From about the 6,000 level onwards, the additional coverage for each 1,000 band of vocabulary is very small indeed, at just a fraction of a percentage point. This makes it impossible to set a frequency level where the coverage falls off in a noticeable way; rather at these lower frequency levels there is a gradual and relatively consistent tailing off. This is obvious if we examine the traditional 10,000 level. The coverage gained at this level (0.32%) is not much different than higher (6,000 = 0.70%) or lower (14,000 = 0.10%) frequency levels. Thus frequency information by itself gives little real help in setting a low-frequency boundary.

(Table 2)

There are two other common ways of conceptualizing low frequency vocabulary. The simplest conceptualization (as a high/low frequency dichotomy) is untenable, as the

vocabulary immediately beyond the 3,000 high-frequency cut-off (i.e. at the 4,000 and 5,000 levels) is clearly too useful to be written off as low-frequency vocabulary. The other is related to the selection of vocabulary in pedagogic materials. Here vocabulary is commonly conceptualized as in the graph below:

High-frequency vocabulary (frequent in all discourse)	AWL (frequent in academic discourse)	Low-frequency vocabulary (rare in all discourse)

In this conceptualization[9], academic vocabulary (as exemplified by Coxhead's *Academic Word List (AWL)* (2000, 2011)) is the next 'band' to teach after high-frequency vocabulary, and everything after that is de facto low-frequency vocabulary as it is rarely addressed in any principled manner. A review of textbooks (e.g. Richmond 2007; Beglar and Murray 2009; Smith-Palinkas & Croghan-Ford 2009) aimed at the highest levels of intensive English programs shows that explicit treatment of vocabulary rarely goes beyond the AWL even though students exiting these programs will progress directly into university study where even introductory textbooks require knowledge of vocabulary up to the 9,000 frequency level (Sutarsyah, Nation, & Kennedy 1994).

The AWL was conceived of as academic support vocabulary which exists *beyond* the high-frequency general vocabulary of English, which Coxhead operationalized as the 2,000 word families in the *General Service List (GSL)* (West 1953)[10]. However, it is easy to see that the above tripartite division of vocabulary is not viable when the AWL is

subjected to a *Lextutor BNC-20* frequency analysis. In fact, we find that 64.3% of the AWL headwords are from the first 3,000 most frequent words in English, while the 4,000 level gives 81.5% coverage, and the 5,000 level 92.1% coverage (Cobb 2010). Thus, although high-frequency vocabulary and academic support vocabulary may be considered different conceptual categories of lexis, in reality, the 3,000 word families of high-frequency vocabulary largely subsume the AWL (see also Hancioğlu, Neufeld and Eldridge, 2008), and so low-frequency vocabulary cannot reasonably be defined as the lexis beyond high-frequency+AWL vocabulary despite what we commonly see in pedagogic materials.

Probably the most fruitful method of establishing a general boundary of low-frequency vocabulary is with a usage-based approach. Hazenberg & Hulstijn (1996) analyzed one corpus of contemporary written Dutch and one corpus of academic Dutch in order to determine how much vocabulary was needed to manage university study. They concluded that it took around a minimum of 10,000 base words (essentially word families) to obtain adequate coverage of these corpora. Although Dutch and English are different languages (but closely related), the 10,000 figure began to be cited for English as a figure which would allow advanced language use (e.g. study at university). It was also given credence by Nation's (1990) setting of the most advanced level on his *Vocabulary Levels Test* at the 10,000 level, even though the test preceded Hazenberg and Hulstijn's empirical evidence. The result was that anything beyond 10,000 word families (which enabled advanced use in the Dutch context) came to be accepted as a rather impressionistic boundary for English low-frequency vocabulary.

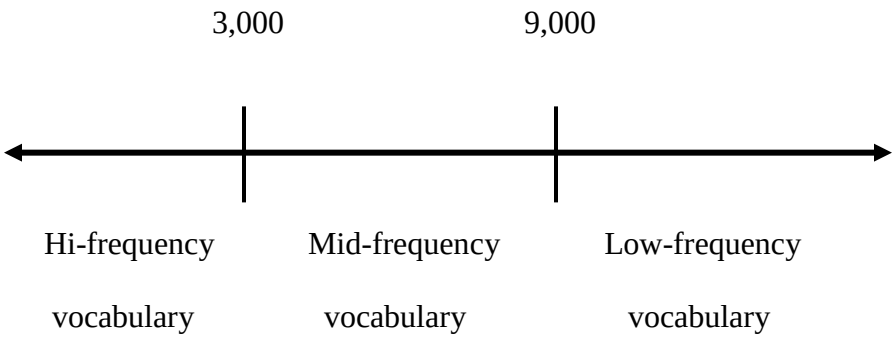
A more recent and relevant empirical study is Nation's (2006) corpus study. He analyzed a range of English authentic texts (novels, newspapers), and calculated that it requires knowledge of the most frequent 8,000-9,000 word families (+proper nouns) to reach the 98% coverage which is the percentage thought to enable efficient reading. It took less vocabulary to cover the spoken corpora at 98% (5,000-6,000 word families). If 8,000-9,000 word families is enough to enable both listening and reading of a wide range of texts without being unduly constrained by a lack of vocabulary knowledge, then low-frequency/utility vocabulary can plausibly be defined as anything beyond this frequency level, i.e. vocabulary beyond the 9,000 frequency band (9,000+).

Nation's 8,000-9,000 word families figure is given support from an analysis of the Corpus of Contemporary American English (COCA) (Davies, 2008). The 425+ million token COCA is a very large corpus of current American English, with a substantial spoken component (for the following analysis, numerals, words with apostrophes, and proper nouns were excluded, leaving 402,646,672 tokens). It is now the best corpus of general English in existence (in terms of size, balance and currency). Using Nation's BNC frequency lists, we find that the most frequent 9,000 word families cover 95.5% of the COCA (Table 3). This means that the most frequent 9,000 word families cover over 95% of a huge amount (400+ million words) of very diverse written and spoken English. The average person would come across much less English than this, and importantly, many fewer different words. Thus the lexical coverage figures would be higher for the amount of language any individual person might be exposed to (Nation, 2001b), and so Nation's (2006) 8,000-9,000 figures are likely to get close to 98% coverage for individual users, especially if numerals and proper nouns are assumed to be known.

Based on this recent corpus evidence, we therefore propose that the low-frequency boundary be moved down from the traditional 10,000+ level to the 9,000+ level. While this may not seem like a large change, the 'savings' to learners is significant if they do not have to master these additional 1,000 word families.

4. Mid-frequency vocabulary

The previous sections have argued that high-frequency vocabulary in English extends up to about 3,000 word families, and that low-frequency vocabulary begins at about the 9,000 frequency level. This leaves a great gap between the 3,000 and 9,000 levels which has not been systematically addressed before. We propose to label this in-between frequency band MID-FREQUENCY vocabulary. It is important that this frequency band is given a name, because it allows the field to recognize it as a discrete phenomenon, with its own unique benefits for users, and pedagogical challenges for language practitioners.



4.1 The nature and benefits of mid-frequency vocabulary

Perhaps the best way of discussing mid-frequency vocabulary is by giving examples and explaining how mid-frequency vocabulary relates to language use. The list below exemplifies the type of words at each 1,000 level in the mid-frequency band:

3,001-4,000: *academic, consist, exploit, rapid, vocabulary*
4,001-5,000: *agricultural, contemporary, dense, insight, particle*
5,001-6,000: *cumulative, default, penguin, rigorous, schoolchildren*
6,001-7,000: *axis, comprehension, peripheral, sinister, taper*
7,001-8,000: *authentic, conversely, latitude, mediation, undergraduate,*
8,001-9000: *anthropology, fruitful, hypothesis, semester, virulent*

It is definitely worth learning mid-frequency words like these, because research demonstrates that accumulating increasing amounts of vocabulary in the mid-frequency range leads to very clear rewards.

One very important reward is the ability to engage with English for authentic purposes, e.g. watching movies. For example, Webb and Rodgers determined that knowing 3,000 word families provides a little over 95% coverage for a range of television programs (2009a) and movies (2009b). This may be enough to enable a reasonable degree of comprehension, but there still would be around 4-5% unknown vocabulary. This translates to about 3.9 unknown words per minute. The authentic purpose for watching television and movies is typically pleasure, and this amount of unknown vocabulary may impact on the learners' ease of viewing, and therefore enjoyment. However, second language listeners who know 98% of the words would face

only 1.6 unknown words per minute, which should enhance the viewing experience. Achieving 98% coverage is largely dependent on mastering words in the mid-frequency range—in movies, around 5,000 word families for horror, drama and crime, and up to 9,000-10,000 families for war and animation. One might expect content-dense television programs such as news broadcasts to require even more vocabulary, and this would be correct: Webb & Rodgers found that it took 4,000 word families to reach 95% coverage and 8,000 families to reach 98%. Because the usual purpose of watching the news is to be informed, it would presumably take nearer the 8,000 figure to fully exploit this information-rich form of communication.

'Authentic purpose' rewards also apply to reading. One very common purpose is to read novels and magazines for pleasure, and this pleasurable reading should not be overly taxing. Carver (1994) explored the relationship between the relative difficulty of written texts and the amount of unknown words in those texts. The study involved two different text types (fictional and factual) and native English primary school and postgraduate university students. He concluded that easy texts generally contained around 0% unknown words, difficult texts around 2% or more unknown words, and texts that were of an appropriate difficulty level around 1% unknown words. This suggests that a 98% coverage figure is none too stringent for pleasure reading, and Nation's (2006) calculations using this figure indicate a vocabulary requirement of 8,000-9,000 word families + proper nouns, again entailing a large amount of mid-frequency vocabulary. Likewise, Nation found that a similar level of vocabulary is necessary to read a range of newspapers.

Another very important authentic purpose is to read English textbooks in English-medium education. For that matter, even university students who are studying for degrees in their L1 are increasingly finding that their subject textbooks are in English: e.g. in Germany, Sweden, Taiwan, and Thailand (Pecorari, et al. 2011). As the purpose is to extract information from these texts, good comprehension is essential. Laufer & Ravenhorst-Kalovski (2010) found that university students in Israel needed enough vocabulary to cover 98% of the examination reading texts (between 6,000 – 8,000 word families), in order to obtain a score on a university-entrance examination which indicated they could read academic material independently (with or without the aid of a dictionary). However, even the ability to read with some guidance and help required 95% coverage, entailing knowledge of between 4,000 and 5,000 word families. Thus, even assisted reading in an educational setting requires a considerable progression into mid-frequency vocabulary.

Two other points are of interest in the Laufer & Ravenhorst-Kalovski study. First, while the students with vocabulary sizes of between 6,000 – 8,000 word families typically achieved exam scores which exempted them from taking an English reading skills class, students with sizes of between 4,000 – 5,000 families typically achieved scores which required one semester of this class, and those with lower sizes required two or three semesters. However, informal reports from both teachers and learners indicated that many of the students with a vocabulary size of around 3,000 families continued to have difficulties with reading even after they had completed the required three semesters of English support classes. So the time and effort that these students spent in learning mid-frequency vocabulary prior to beginning university was paid back when they did not

have to take semesters of English reading classes. Furthermore if they do not have this vocabulary, they may not be able to achieve the necessary levels for reading university academic texts, even with the help of supplementary reading classes.

The second point is that improvement in reading test scores was closely connected with progression through the mid-frequency vocabulary. An increase of vocabulary from the 4,000 to 5,000 frequency levels increased reading scores just as much as an increase from the 3,000 to 4,000 levels. In fact, the best improvement in the reading scores came from vocabulary increases from the 5,000 – 6,000 and 5,000 – 7,000 levels. Thus, even though the percentage of text coverage decreases as one moves through mid-frequency vocabulary (e.g. 2.2% from 3,000 – 4,000 vs. 1.3% from 5,000 – 7,000), the later stages of mid-frequency vocabulary seem just as, if not more, important for effective reading.

A different kind of reward relates to the fluency with which a learner can use their vocabulary. Laufer & Nation (2001) looked at the relationship between vocabulary size on the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham 2001), and the speed at which learners could answer items on that test. They found that increased speeds on higher-frequency 3,000 and AWL sections began only when learners reached a vocabulary size of around 5,000 word families. Furthermore, the more vocabulary known beyond this level, the faster the speed, with the size/speed relationship strongest at the 10,000 frequency level (r=.67). Thus, knowledge of vocabulary into the mid-frequency levels corresponds with not only knowledge of that lexis, but also improved speed of access for both mid- and high-frequency words. While increased speeds in answering a vocabulary test is not the same as accessing vocabulary in the four skills, it is suggestive. A lack of fluency can have a major impact on the way English can be used, even by

highly-proficient learners. McMillion & Shaw (2008) contrasted Swedish and British university biology students reading English texts and concluded that their advanced Swedish learners of English could reach virtually the same comprehension levels as the British students. However, the Swedish students consistently read at rates 25% slower than the British students. This means these students may be disadvantaged in two ways. First, they need to spend 25% more time reading in order to reach comprehension levels on par with L1 readers. Second, when this time is not available (e.g. under exam conditions), they will not be able to demonstrate comparable levels of comprehension.

Our discussion of mid-frequency vocabulary highlights its importance for operating in English across a range of topics and situations. But what of learners who are specializing in one area; can they make do with specialized English, e.g. Business or Medical English? Lists of technical vocabulary have been promoted as a way of focusing the vocabulary study in such specific domains (e.g. Hyland and Tse, 2007). These lists vary widely in both their scope and how much coverage they provide of the specialized texts in the target domain (e.g. 113 word families with 3.7% coverage of theology lectures (Lessard-Clouston, 2010); 623 word families for 12.24% coverage of medical research articles (Wang, Liang, & Ge 2008) and 2,000 word families for 95% coverage of foundation level engineering textbooks (Ward, 1999)). We agree that using such lists can be a useful aid in determining which of the mid-frequency words to focus on first, but it is important to realize that that high-frequency+technical words are not enough to cope with domain-specific texts; that is, mid-frequency vocabulary is still required. There are a number of reasons for this:

1. Text coverage of high- + academic + technical vocabulary often does not reach 95-98% (e.g. Chung & Nation, 2003; Fraser, 2005), and so knowledge of mid-frequency vocabulary may be necessary to reach these coverage levels.[11]
2. While a number of technical words have very specialized meanings and are low-frequency, many of them have more generalized meanings and come from the high- and mid-frequency bands. Thus, learners who know high- and mid-frequency vocabulary have a headstart when learning lists of technical vocabulary.
3. Technical words are often defined in the text, but one must know the surrounding words (high- and mid-frequency) in order to understand the definitions.
4. The compilers of technical lists normally take a very narrow approach to defining learners' needs, e.g. being able to read engineering textbooks or understanding theology lectures, which does not take into account possible wider or longer-term needs, e.g. speaking English in the workplace or reading the newspaper. Mid-frequency vocabulary is necessary to participate in this wider range of activities.

4.2 The lack of a principled approach to teaching mid-frequency vocabulary

We have seen the benefits of developing a relatively large vocabulary, but the three different frequency bands (high, mid, low) have been treated quite differently in teaching. High-frequency vocabulary is already addressed to some extent by teaching pedagogy, as textbooks, word lists, graded readers, and learner dictionaries all focus on this vocabulary. Additionally, the high frequency means that learners will be relatively well-exposed to this vocabulary in any input they receive. Unfortunately, many learners still

do not master high-frequency vocabulary, even after 1,000 hours or more of English instruction (Laufer, 2000). We suggest that, as a minimum, English language programs emphasize teaching high-frequency vocabulary up to the 3,000 frequency level.

On the other end of the frequency continuum, low-frequency vocabulary is not typically useful enough to warrant an explicit focus, and Nation (1990) has long argued that it should be left to learners to deal with it themselves through the use of learning strategies. This seems sensible, but despite this, the topic-based focus of many materials means that low-frequency vocabulary regularly gets explicit attention because it is seen to be necessary for the comprehension of particular reading or listening texts. It would be useful for materials writers to either gloss this vocabulary and/or use text-profiling tools (e.g. *Lextutor*) to minimize the low-frequency vocabulary and replace it where possible with either high-frequency vocabulary (if the task purpose is fluency practice) or mid-frequency vocabulary (if the purpose includes learning new words) (Nation 2009).

This leaves mid-frequency vocabulary, which is much more problematic. It is not often addressed pedagogically, yet we have seen its considerable importance and benefits. We thus have a situation where vocabulary needed by learners is not addressed in any principled way. Some teachers might assume that vocabulary will somehow be 'picked up' from exposure to various language activities within the classroom and from natural input outside the classroom.

Unfortunately, there is some evidence that mid-frequency vocabulary is not typically used or taught in classrooms by teachers to any great extent. Perhaps unsurprisingly, Horst, Collins, & Cardoso (2009) found that the vast majority of cases of direct vocabulary teaching in primary ESL classrooms (Grade 6) focused on high-

frequency vocabulary, with very little focus on mid-frequency vocabulary. However, there is not necessarily a greater emphasis on mid-frequency vocabulary at later stages of language learning. Tang & Nesi (2003) studied the teacher talk of two secondary school teachers in Guanzhou and Hong Kong and found that only 6 and 12% respectively of their vocabulary went beyond 3,000 word families (in these cases, the first 2,000 + a 1,000-item list made up of words from secondary school and university texts). Horst (2010) analyzed 32 hours of classroom discourse from a high-intermediate/advanced adult ESL class. Of the 121,967 words of teacher talk, 118,330 (97%) were high-frequency vocabulary, and only 2,521 (2%) came from the mid-frequency band. Furthermore, there was generally not enough repetition of these words to facilitate acquisition. Thus, across a variety of teaching contexts, the opportunities for learning mid-frequency vocabulary from teacher talk remain surprisingly low. This conclusion is supported by Folse's (2010) finding that not only are cases of explicit vocabulary instruction relatively rare, but when they do occur they are usually not done in a way that facilitates remembering or recycling, e.g. given orally with no accompanying visual cues, or without drawing the whole class's attention to the word.

Similarly, mid-frequency vocabulary does not seem to be systematically addressed in textbooks either. Matsuoka & Hirsh (2010) analyzed the vocabulary from the best-selling *New Headway Upper-intermediate* English textbook and found that high frequency vocabulary (GSL + AWL + proper nouns + 32 other word families that were assumed to be known) provided 95.5% coverage of the textbook's 44,877 running words. Of the 1,005 remaining word families, 66.4% occurred only once and only 12.1% occurred 5 times or more. While textbooks are typically used under teacher guidance,

which may lead to more noticing and engagement with the target vocabulary than might be the case with unassisted reading, these figures are still not promising. The authors of the series state that the books contain a "very strong lexical syllabus" (Soars & Soars, 1996: v), but Matsuoka & Hirsh's results show that this upper-intermediate textbook provides few opportunities for learning words at mid-frequency or beyond. But what of the other levels? We submitted the single words from the wordlist in *New Headway Intermediate* to a *Lextutor BNC-20* frequency analysis and found that it includes 440 word families, of which only 110 come from the mid-frequency band. The wordlist from *New Headway Advanced* includes 782 families, with only 427 mid-frequency families. Given the vocabulary requirements outlined in this paper, both the total number of target words, and the number of mid-frequency words seem rather small. While we do not know how much recycling of mid-frequency vocabulary there is throughout these two books, the small amount of recycling in *New Headway Upper-intermediate* suggests that it is probably not enough to reliably promote acquisition, unless teachers take up this particular vocabulary for active instruction in the classroom. Furthermore, even if there is recycling across the levels in the series, the length of time required to get through even one level means that the time between meetings is too long.

Hsu (2009) examined the 20 international General English (GE) textbooks used at her university in Taiwan (ranging from low-intermediate to advanced) in order to determine how much vocabulary was required to achieve 95% coverage of the reading passages. The main articles in each book were analyzed for word frequency using Nation and Heatley's (2002) RANGE program with the BNC lists. Her findings show little uniformity between the level of the textbook and the vocabulary required both within and

across textbook series (Table 4). This study illustrates the lack of a standardized approach to vocabulary in language textbooks, particularly in relation to reading difficulty, with materials writers seemingly unaware of vocabulary grading (through frequency) to consistently aid reading comprehension and develop vocabulary through the textbook levels. For example, the advanced *Reading for Real* required 4,000-4,500 word families to reach 95%, while the low-intermediate *Reading for Success 2* required 7,000-7,500 families. Hsu reports that the Taiwanese high school curriculum covers 2,000 words. 95% text coverage can be considered an appropriate instructional level (leaving 5% of the words available to be learned) for learners aiming to become independent readers (98%+ coverage). However, few of the books in this study offer optimum learning conditions for increasing learners' vocabulary size or improving their reading ability. Clearly there needs to be more consistency across textbook series, but this can only happen if vocabulary grading becomes a primary consideration of textbook writers. Hsu's figures clearly show the importance of mid-frequency vocabulary for reading, because for every textbook except *Select Readings Intermediate*, substantial amounts of mid-frequency vocabulary is necessary to get to 95% coverage.

The studies reviewed in this section clearly show that mid-frequency vocabulary is necessary for a wide range of language uses, but also that neither teacher talk nor textbooks appear to address it in a principled manner. This raises a number of pedagogic issues, some of which we will consider in the next section.

5. Research agenda for mid-frequency vocabulary

Mid-frequency vocabulary poses a serious pedagogic challenge in how to deal with the thousands of word families in the band. We feel that explorations in the following areas would go some way towards providing insight into how to address this challenge.

● What is the total vocabulary input when both teacher input and materials input are combined? Research to date has tended to focused on one or the other.

● At what rates can we reasonably expect learners to acquire vocabulary? Milton (2009: 89) surveys a range of studies and concludes that "learners, as a very general average, appear to gain about four words per hour from regular classroom contact." Is this rate a cognitive learning constraint or an artefact of an insufficient focus on vocabulary?

● It takes words to learn words. Many learning strategies rely on knowledge of high-frequency vocabulary (e.g. using dictionaries, keeping vocabulary notebooks). Is it possible for language programs to set out more ambitious early vocabulary targets and achieve them through a 'vocabulary flood' of the 3,000 high-frequency words?

● To what degree is it feasible to manipulate the occurrences of mid-frequency vocabulary in learning materials to enable sufficient recycling to occur? Is it only possible to do this with computer-based materials or can it be done in traditional textbooks?

● Is it possible to develop a series of more advanced graded readers in which mid-frequency vocabulary is supported through techniques such as glossing or elaboration in the text (e.g. Nation, 2009)?

● To what extent can computerized vocabulary learning programs contribute towards learners' ability to use vocabulary in communicative contexts?

● Should a standard vocabulary size be attached to different textbook levels (e.g. lower intermediate, advanced), so that textbooks can be more comparable across series, and to ensure lexical progression within a series?

6. Conclusion

The main purpose of this paper has been to provide workable, empirically-based definitions of high-, mid-, and low-frequency bands, and to highlight mid-frequency vocabulary so that it can be discussed as a phenomenon in its own right. We have highlighted a number of areas which require further research to determine how mid-frequency vocabulary should be addressed pedagogically. We hope that the concept of mid-frequency vocabulary will lead to more realistic vocabulary size targets in language programs and learner materials and classroom research into their effectiveness.

Notes

1. The ideas in this paper were developed jointly by the two authors. A preliminary conceptualization of the ideas was jointly presented at AAAL 2011, and a revised version was presented as a plenary talk at Alberta TESL 2011 by the first author. This paper is a slightly revised version of the plenary talk, with improvements suggested by five reviewers.

2. A serious limitation of the discussion in this paper is that it is based around individual word forms/families, and does not take account of the ubiquitous nature of formulaic

language. This is because most vocabulary research to date has only counted individual word forms/families. See Simpson-Vlach & Ellis (2010) and Martinez & Schmitt (under review) for two phrasal lists which aim to address this deficiency.

3. A *word family* includes a root form (*select*), its inflections (*selected, selecting, selects*), and its derivatives (*selection, selective, selectively, preselect).* It should be noted that the vocabulary size figures for individual word forms (e.g. *select*, *selecting*, and *selective* all treated as separate words) would be far higher than the word family figures in this paper. For, example, it has been estimated that 8,000 families (enabling wide reading, Nation 2006) entail 34,660 individual words.

4. Nation based his 1,000 frequency bands on the BNC, which contains mainly written British and Irish English. See his 2006 article and his website <http://www.victoria.ac.nz/lals/staff/paul-nation.aspx> for the details of his methodology and its limitations.

5. One reason the 1st 1,000 level has so much coverage is that function words are very frequent and cover so much text just by themselves. For example, function words make up 43% of the written and spoken English in the COCA (Mark Davies, personal communication).

6. The current consensus is that 98% lexical coverage (i.e. the percentage of words known in a written text) is necessary for adequate comprehension, that is, only two

unknown words per 100 (Hu & Nation 2000; Laufer & Ravenhorst-Kalovski 2010; Schmitt, Jiang, & Grabe, 2011). 95% coverage is workable, but less than ideal. Of course, knowing these amounts of vocabulary does not guarantee reading comprehension, as reading involves more than just vocabulary knowledge. However, research indicates that if readers know enough words to cover 95%-98% or more of a text, they are likely to obtain 60% - 68% comprehension of that text (Schmitt, et al., 2011).

7. Participants in this study achieved about 75% comprehension of the listening passages at the 95% lexical coverage rate, compared to 96% comprehension at 100% coverage. Staehr (2009) found evidence that advanced listening (using the Certificate of Proficiency in English (CPE) listening test) requires 98% coverage of the passages.

8. CANCODE is the Cambridge and Nottingham Corpus of Discourse in English, a 5-million word corpus of unscripted spoken English.

9. This conceptualization has been partially driven by research done with Nation's early *VocabProfiler*, which essentially breaks vocabulary into only three categories: 1st and 2nd 1,000 vocabulary (high-frequency), Academic vocabulary, and Off List (all other words).

10. Although use of the GSL subsequently became controversial (Coxhead, 2011), at the time of compilation, prior to 1998, it was the best corpus resource available.

11. Our 3,000 definition of high-frequency vocabulary includes most AWL words. However, much of the technical word list research uses Nation's four categories, in which academic vocabulary is separated from the most frequent 2,000 word families.

Acknowledgements

We are extremely grateful to Mark Davies for supplying the COCA frequency and coverage figures in Endnote 5 and Table 3, and to Tom Cobb and Marlise Horst for their insightful input during the drafting process.

References

Adolphs, S. & N. Schmitt. (2003). Lexical coverage of spoken discourse. *Applied Linguistics* 24.4, 425–438.

Beglar, D. & N. Murray. (2009). *Contemporary topics 3: Academic listening and note-taking skills* (3rd edn). Harlow: Pearson.

Carver, R.P. (1994). Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Reading Behavior* 26.4, 413–437.

Chung, T.M. & P. Nation. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language*, 15.2, 103–116.

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology* 11.3, 38–63.

Cobb, T. (2010). Learning about language and learners from computer programs. *Reading in a Foreign Language* 22.1, 181–200.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly* 34.2, 213–238.

Coxhead, A. (2011). The Academic Word List 10 years on: Research and teaching implications. *TESOL Quarterly* 45.2, 355–362.

Davies, M. (2008-). *The Corpus of Contemporary American English: 425 million words, 1990-present*. Available online at <http://corpus.byu.edu/coca/>.

Folse, K. (2010). Is explicit vocabulary focus the reading teacher's job? *Reading in a Foreign Language* 22.1, 139–160.

Fraser, S. (2005). The lexical characteristics of specialized texts. In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (eds.), *JALT2004 conference proceedings*. Tokyo: JALT.

Hancioğlu, N., S. Neufeld, & J. Eldridge. (2008). Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes* 27, 459–479

Hazenberg, S. & J. H. Hulstijn. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics* 17.2, 145–163.

Horst, M. (2010). How well does teacher talk support incidental vocabulary acquisition? *Reading in a Foreign Language* 22.1, 161–180.

Horst, M., L. Collins, & W. Cardoso. (2009, March). Focus on vocabulary in ESL teacher talk. Paper presented at the annual conference of the American Association for Applied Linguistics, Denver, CO.

Hsu, W. (2009). College English textbooks for general purposes: A corpus-based analysis of lexical coverage. *Electronic Journal of Foreign Language Teaching* 6.1, 42–62.

Hu, M. & I. S. P. Nation. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language* 23.1, 403–430.

Hyland, K. & P. Tse. (2007). Is there an ‘‘academic vocabulary”? *TESOL Quarterly*, 41.2, 235–253.

Laufer, B. (2000). Task effect on instructed vocabulary learning: The hypothesis of 'involvement'. *Selected Papers from AILA ’99 Tokyo* (pp. 47–62). Tokyo: Waseda University Press.

Laufer, B. & P. Nation. (2001). Passive vocabulary size and speed of meaning recognition: Are they related? *EUROSLA Yearbook* 1, 7–28.

Laufer, B., & G. C. Ravenhorst-Kalovski. (2010). Lexical threshold revisited: Lexical coverage, learners’ vocabulary size and reading comprehension. *Reading in a Foreign Language* 22, 15–30.

Lessard-Clouston, M. (2010). Theology lectures as lexical environments: A case study of technical vocabulary use. *Journal of English for Academic Purposes*, 9, 308–321.

Martinez, R. and N. Schmitt. (under review). A Phrasal Expressions list.

Matsuoka, W. & D. Hirsh. (2010). Vocabulary learning through reading: Does an ELT course book provide good opportunities? *Reading in a Foreign Language* 22.1, 56–70.

McMillion, A. & P. Shaw. (2008). The balance of speed and accuracy in advanced l2 reading comprehension. *Nordic Journal of English Studies* 7.3, 123–143.

Nation, I.S.P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

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

Nation, I.S.P. (2001b). How many high frequency words are there in English? In M. Gill, A.W. Johnson, L.M. Koski, R.D. Sell, and B. Wårvik (eds.) *Language, Learning and Literature: Studies Presented to Håkan Ringbom.* Åbo Akademi University, Åbo: English Department Publications 4. pp. 167-181.

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. (2009). New roles for L2 vocabulary? In L. Wei and V. Cook (eds.) *Contemporary Applied Linguistics Volume 1: Language Teaching and Learning* London: Continuum. pp. 99–116.

Nation, I.S.P. (2011). Research in practice: Vocabulary. *Language Teaching* 44.4, 529–539.

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

Pecorari, D, P. Shaw, H. Malmström, & A. Irvine. (2011). English textbooks in parallel-language tertiary education. *TESOL Quarterly* 45.2, 313–333.

Read, J. (2000). *Assessing Vocabulary.* Cambridge: Cambridge University Press.

Richmond, K. (2007). *Inside reading 4: The Academic Word List in context.* New York: Oxford University Press.

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

Schmitt, N., D. Schmitt, & C. Clapham. (2001). Developing and exploring the behaviour of two new versions of the Vocabvulary Levels Test. *Language Testing* 18.1, 55–88.

Schmitt, N, X. Jiang, & W. Grabe. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal* 95.1, 26–43.

Schonell, F.J., I. G. Meddleton, & B. A. Shaw. (1956). *A study of the oral vocabulary of* adults. Brisbane: University of Queensland Press.

Simpson-Vlach, R. & N. C. Ellis. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31.4, 487–512.

Smith-Palinkas, B. & K. Croghan-Ford. (2009). *Key concepts: Reading and writing across the disciplines*. Boston, MA: Heinle Cengage Learning.

Soars, L & J. Soars. (1996). *New Headway Intermediate: Teacher's Book*. Oxford: Oxford University Press.

Staehr, L. (2009). Vocabulary knowledge and advanced listening comprehension in English as a Foreign Language. *Studies in Second Language Acquisition* 31, 577–607.

Sutarsyah, C., P. Nation, & G. Kennedy. (1994). How useful is EAP vocabulary for ESP? A corpus based case study. *RELC Journal* 25, 34–50.

Tang, E. & H. Nesi. (2003). Teaching vocabulary in two Chinese classrooms: Schoolchildren's exposure to English words in Hong Kong and Guangzhou. *Language Teaching Research* 7.1, 65–97.

Thornbury, S. (2002). *How to teach vocabulary.* Harlow: Longman.

van Zeeland, H. & N. Schmitt. (under review). Lexical coverage and L1 and L2 listening comprehension: The same or different from reading comprehension?"

Wang, J., S-L. Liang, & G-C. Ge. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, 27, 442–458.

Ward, J.W. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language*, 12.2, 309–324.

Webb, S. & M. P. H. Rodgers. (2009a). Vocabulary demands of television programs. *Language Learning* 59, 335–366.

Webb, S. & M. P. H. Rodgers. (2009b). The lexical coverage of movies. *Applied Linguistics* 30, 407–427.

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