How Much Vocabulary is Needed to Use English? Replication of Van Zeeland & Schmitt (2012), Nation, (2006), and Cobb (2007)

Norbert Schmitt

University of Nottingham

Tom Cobb

University of Quebec at Montreal

Marlise Horst

Concordia University

Diane Schmitt

Nottingham Trent University

Authors' bios

Norbert Schmitt is Professor of Applied Linguistics at the University of Nottingham. He is interested in all aspects of second language vocabulary. He has published 8 books (with the latest being *Researching Vocabulary: A Vocabulary Research Manual* (2010, Palgrave Macmillan)), 50 journal articles, and 23 book chapters on various vocabulary topics. He currently sits on the editorial board of *Language Testing*. His personal website (www.norbertschmitt.co.uk) gives much more information about his research, and also

provides a wealth of vocabulary resources for research and teaching.

norbert.Schmitt@nottingham.ac.uk

Tom Cobb is professor of Didactique des langues (or Applied Linguistics) at the University of Quebec at Montreal. He too is interested in all aspects of second language vocabulary, but has focused mainly on those that can be investigated or learned with the help of a computer. His Compleat Lexical Tutor website (www.lextutor.ca) has a host of resources for learners, teachers, and researchers as well as links to his research studies.

cobb.tom@uqam.ca

Marlise Horst is an Associate Professor in the Department of Education at Concordia University in Montreal, where she teaches courses in second language vocabulary acquisition and the history of English for language teachers. Her current research explores opportunities to learn new L2 vocabulary via exposure to classroom input. She is an avid quilter in her spare time. marlise@education.concordia.ca

Diane Schmitt is a Senior Lecturer in EFL/TESOL at Nottingham Trent University and Chair of BALEAP. She teaches on the MA in English Language Teaching and also on a range of English for Academic Purposes courses. She has co-authored two textbooks on teaching vocabulary. Her areas of interest include: academic writing, plagiarism, vocabulary acquisition, language testing, materials development and the international student experience. diane.schmitt@ntu.ac.uk

Abstract

There is current research consensus that L2 learners are able to adequately comprehend general English written texts if they know 98% of the words that occur in the materials. This important finding prompts an important question: How much English vocabulary do ESL learners need to know to achieve this crucial level of known-word coverage? A landmark paper by Nation (2006) provides a rather daunting answer. His exploration of the 98% figure with a variety of spoken and written corpora showed that knowledge of around 8,000-9,000 word families is needed for reading and 6,000-7,000 for listening. But is this the definitive picture? A recent study by van Zeeland ~~and~~ & Schmitt (2012) suggests that 95% coverage may be sufficient for listening comprehension, and that this can be reached with the much more manageable figure of 2,000-3,000 word families. Getting these figures right for a variety of text modalities, genres and conditions of reading and listening is essential. Teachers and learners need to be able to set goals, and as Cobb's study of learning opportunities (2007) has shown, coverage percentages and their associated

vocabulary knowledge requirements have important implications for the acquisition of new word knowledge through exposure to comprehensible L2 input. This article proposes approximate replications of Nation (2006), Van Zeeland & Schmitt (2012), and Cobb (2007), in order to clarify these key coverage and size figures.

1. Introduction

People use language to communicate and express meaning, and this meaning is essentially conveyed by vocabulary. Thus knowledge of vocabulary is fundamental to all language use, and so must be learned in some manner in order for learners to become communicative in a new language. However, the lexicons of most languages are very large. For example, Goulden, Nation & Read (1990) estimated there are 54,000 word families in English. Given that most word families have several members (*stimulate, stimulated, stimulating, stimulates, stimulation, stimulative*), this translates to many hundreds of thousands of individual word forms.[1] Even very proficient speakers will not know all of these words,[2] and Goulden et al. found that their New Zealand university undergraduates had an English vocabulary size of about 17,000 word families. This is still far out of reach for most second language learners, and it is not surprising that second language teachers and textbook writers struggle with the sheer number of words that could be taught. What is needed for pedagogical purposes is descriptions of the amount of vocabulary which is necessary, not to match native speakers, but to be functional in specific communicative contexts.

In order to generate such descriptions, two things are required. First, one must know what percentage of the vocabulary in a stretch of spoken or written discourse needs to be known by a learner in order for him or her to understand the discourse. This is known as LEXICAL COVERAGE. People can usually understand speech or writing even if there are a few unknown words, so 100% coverage is not typically necessary. But if too many words are unknown, comprehension is compromised and listening or reading becomes a chore. What percentage of words should be known? Most research suggests that coverage in the range of 95%-98% is adequate for acceptable comprehension, or in other words, that acceptable comprehension can be achieved with 2%-5% of the words unknown (e.g. Hu &

Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Van Zeeland & Schmitt, 2012). Second, when a coverage figure is established, one must determine how many specific words this corresponds to. For example, a typical finding is that 98% coverage in written texts corresponds to knowledge of about 8,000 word families (e.g. Nation, 2006). We will refer to the number of words needed to meet a lexical coverage percentage in various communicative contexts as VOCABULARY SIZE.

Another factor is pedagogical practicality. Since the number of words that can be taught explicitly in language classes is limited, studies would do well to include an empirical consideration of learners' capacity to acquire new vocabulary incidentally, through exposure to reading and listening input. Cobb (2007), in a study using corpus coverage to calculate learning opportunities, showed that learners' typically small vocabulary sizes of 2,500-3,000 word families can be partially explained by the very low rate at which words at subsequent frequency levels (i.e. 3,000+) occur in texts. Words must normally be met a certain number of times (the figure 10 is often cited), but Cobb argues that this number of repetitions is typically not available for less frequent words. If his analysis is correct, then the notion that written language comprehension depends on lexicons on the scale of 8,000 word families amounts to a discouraging picture for learners, since relatively few of them will arrive at that figure. However, learners may well be able to cope with lower levels of lexical coverage/vocabulary size than the research suggests. They may be able to do this through using resources like dictionaries and various forms of online support. If so, the coverage/size figures may be set higher than necessary for real-world use, and new 'resource aided' figures need to be developed.

It is crucial to have good estimates of the vocabulary sizes necessary to be functional in specific contexts and uses of a language, because these estimates form

learning targets for language students. An estimate that is too low could lead to a lowering of pedagogical goals such that learners would not acquire a vocabulary large enough to make competent language use possible. An estimate that is too high would be unnecessarily demotivating for learners, and may include words that are so infrequent that they have little practical utility in normal language use. There are a limited number of studies informing these essential size targets, and so it is vital to replicate and expand upon the ones we have. This paper will suggest replication of studies of lexical coverage (Van Zeeland & Schmitt~~,~~ 2012), vocabulary size (Nation~~,~~ 2006), and plausible learnability (Cobb~~,~~ 2007) in order to develop a more reliable, nuanced, and ecologically valid understanding of the amount of vocabulary learners need to acquire in order to become proficient language users in their chosen domains.

2. The original studies and suggested approaches to replication

2.1 Lexical coverage of spoken discourse (Van Zeeland & Schmitt~~,~~ 2012)

Most research on lexical coverage in relation to L2 comprehension has been conducted on reading, and we now have a fairly good idea of the percentage of vocabulary that needs to be known to allow comprehension of written text. The earliest research in this area indicated that 95% lexical coverage was needed (Laufer~~,~~ 1989). In real terms, 5% unknown vocabulary equates to about one unknown word in roughly every two lines of text, and so over 15 unknown words on every page. Thus, it is perhaps not surprising that subsequent research by Hu ~~and~~ & Nation (2000) suggested a higher coverage figure closer

to 98%. More recently, Schmitt, Jiang~~, and~~& Grabe (2011) investigated each percentage point of coverage between 90 and 100 percent, in an attempt to describe the overall relationship between coverage and comprehension. This revealed a linear relationship between the two, which suggests that the coverage level required depends on the degree of comprehension aimed for. Based on their data, if 60% comprehension is the goal, 98% lexical coverage is needed. Laufer ~~and~~ & Ravenhorst-Kalovski (2010) support the idea of basing the required coverage level on the reading comprehension wished for (e.g. if a learner's goal is to read a second language novel for pleasure, then 100% comprehension may not be needed or worth the learning investment). These authors suggest two lexical coverage thresholds, depending on the definition of 'adequate' comprehension: 98% as the sufficient and 95% as the minimal. Based on the performance of their Israeli participants, the authors conclude that 95% coverage enables acceptable comprehension and is probably viable with some support (e.g. teacher or learner resources like dictionaries) and that 98% coverage leads to successful comprehension by most learners, and is likely to enable independent reading. Overall, the consensus is that about 98% is the lexical coverage which is most appropriate for most purposes involving written text.[3]

In contrast, there has been very little research into the lexical coverage required for listening. The main study to date has been Van Zeeland ~~and~~ & Schmitt (2012). They had their ESL (mixed L1) participants listen to four anecdotes told in the first person about people getting into unusual situations. The stories had various percentages of words replaced with nonwords (0, 2, 5, and 10%), so that percentages of known vocabulary in the stories were precisely 100, 98, 95, and 90% respectively. Participants' comprehension was measured by a 10-item multiple choice test for each anecdote. The researchers found that the participants who knew 100% of the words in a story had a mean score of 9.62, those knowing 98% of the words scored 8.22, those with 95% had 7.65, and those with 90%

knowledge scored 7.35. Overall, knowledge of greater percentages of the vocabulary in the stories led to better test scores, and thus comprehension. However, even at the 95% and 90% knowledge levels, the comprehension was still quite good in absolute terms and good enough for many practical purposes. There was no statistical difference in the test scores between the 95% and 90% knowledge levels, but the standard deviations were large at the 90% level, indicating that there was real variability in learners' ability to cope with this low amount of lexical coverage. Van Zeeland ~~and~~ & Schmitt thus concluded that 95% lexical coverage was the more reasonable criterion for adequate comprehension, because at this level, the performances were much more consistent among the participants.

The Van Zeeland ~~and~~ & Schmitt study is a good start, but replications could usefully address its inevitable limitations. We suggest *approximate replications* of this study (Porte 2012), where certain variables might be changed to determine how generalizable the original paper's results are, and to either strengthen or challenge the conclusions of that paper. Van Zeeland ~~and~~ & Schmitt used informal narratives of about two minutes length in their study, with repeated listenings, and they acknowledge that their results might be viewed as 'best-case' performance. The most obvious variable to start exploring is the type of listening. Narratives typically have a straightforward chronological structure, which should make listening easier than, say, a lecture or a detailed explanation. This is especially true because listeners rely more on top-down processing than readers (Lund~~,~~ 1991; Park~~,~~ 2004). This suggests that listening comprehension may be largely based on factors such as world knowledge and topic familiarity. Such top-down information is believed to be compensatory in use, in the sense that it can be employed strategically by listeners to compensate for inadequate knowledge of the L2 or an inability to recognize words in continuous speech (Field~~,~~ 2004; Vandergrift~~,~~ 2011). Thus passages with more obvious organization (such as narratives) should be easier to comprehend when listening,

and indeed narratives have been found to be most comprehensible genre for listeners (Rubin~~,~~ 1994). It is an open question whether types of discourse with a less obvious organization (e.g. everyday chat, jokes, political speeches) can also be comprehended with 95% coverage, and these types should be explored. It is not always obvious *a priori* whether they might require higher or lower lexical percentages for comprehension. For example, everyday chat, with its numerous digressions and topic changes, might require a higher lexical coverage for comprehension. Conversely, the greater opportunities for questions and clarification might allow comprehension with a lower percentage of coverage.

Another variable which could be usefully explored is the length of listening. Van Zeeland ~~and~~ & Schmitt's passages were relatively long in experimental terms at about two minutes. However, many real-world listening contexts, such as attending to academic lectures, political speeches, and radio talk shows require much longer periods of concentration. It would be interesting to determine whether truly extended listening contexts (20+ minutes) would be comprehensible with 95% coverage, or whether the unknown words would eventually begin to affect comprehension, or at least make listening onerous. On the other hand, it might be that as the sense of the message begins to accumulate, more top-down processing can come into play and listening becomes easier.

Another decision made by the researchers was to allow the participants to listen to the passages twice, in order to avoid memory affecting the comprehension results. However, as most listening is a one-off affair, it would be interesting to know whether single listenings would also be consistently comprehensible at 95% coverage. A straightforward way of assessing this would be to repeat the van Zeeland ~~and~~ & Schmitt study, but have some participants listen to the passages twice, as in the original study, and others only once. If the single listeners do much worse, then this might indicate that the

95% coverage figure is too optimistic, and the coverage figures from the single listenings might be a more appropriate indication of the necessary lexical coverage required for non-interactive listening.

Finally, the researchers decided to situate participants in a condition of not being able to ask questions. This is convenient for research (as a way of equalizing the participants, since some will ask more questions than others due to personality factors) but it comes at some cost in ecological validity. Only a small proportion of real-life interpersonal listening takes place with no option to interact or ask questions (although this clearly is the case with media exposure such as listening to radio, TV, movies, online lectures, YouTube videos, etc). In any case, the study should be replicated under more typical interpersonal conditions, which could be predicted to lower the coverage needed to a point below 95%.

2.2 Vocabulary size (Nation~~,~~ 2006)

Perhaps the most important vocabulary size figures to establish definitively are those relating to L2 learners' ability to be functional in general English in both the written and spoken modes. Once established, these figures should inform all non-specialized (e.g. non-ESP) English teaching pedagogy and materials design. To date the most influential paper in this area is undoubtedly Nation's 2006 study. Using a mini-corpus of five English novels (*Lord Jim, Lady Chatterley's Lover, The Turn of the Screw, The Great Gatsby,* and *Tono-Bungay*), Nation calculated that a learner would be required to know about 4,000 of the most frequent word families plus proper nouns to reach 95% lexical coverage, and around 8,000-9,000 families plus proper nouns to reach 98% coverage. He found similar figures for

a corpus of newspapers. Turning to unscripted spoken English, Nation used two parts of the Wellington Corpus of Spoken English (n.d.). One part included talk-back radio, where listeners phone in with their spontaneous comments on the issue being discussed, and the other was made up of friendly conversation between family members and friends. Nation found that about 3,000 word families plus proper nouns provided more than 95% coverage, but that it took 6,000-7,000 word families to reach 98% coverage. He also investigated the movie *Shrek*, in which it took 4,000 word families plus proper nouns to reach the 95% lexical coverage, and 7,000 to reach the 98% level. These figures are broadly in line with Webb ~~and~~ & Rodgers' findings based on the scripted talk in movies (2009a) and television (2009b).[4]

To be clear, this research did not involve actual learners with knowledge of the 4,000 or 8,000 highest frequency word families; nor did it involve measuring learners' comprehension. Instead, it used frequency profiling of his target text collections and corpora to determine the number of word families that learners would hypothetically need to know to achieve a particular level of known word coverage (e.g. 98%). Nevertheless, his vocabulary size figures (based on 98% coverage) of 6,000-7,000 word families for spoken English and 8,000-9,000 for written English are very widely cited. Given the impact of this study, it is important to replicate it to confirm (or revise) those figures.

An approximate replication approach also seems appropriate to address Nation's study as there are a number of variables that could usefully be manipulated. For initial replications, we propose to leave unchanged Nation's methodology for deriving vocabulary size, which uses his BNC-based word family lists as counting units, as it is well established and has proven its usefulness. What is needed are replications of Nation that use this same methodology but test much larger corpora of general English. Nation actually used rather small data sets in his influential study: the single novel *Lady Chatterley's Lover* (121,000

words), the five novels mentioned above taken together (474,000 words), a newspaper corpus (440,000 words), the script of the movie *Shrek* (10,000 words), and two parts of the Wellington Corpus of Spoken English (around 200,000 words). Nation's purpose was to determine the vocabulary sizes necessary to read and listen to general English in various contexts of use, so testing these various small sample corpora made sense. However, he conflated the individual results in order to come up with the overall vocabulary size figures discussed above. It is these global figures (6,000-7,000 word families for listening; 8,000-9,000 word families for reading) which now need to be checked with larger, more comprehensive corpora.

Two large current corpora against which the coverages of Nation's lists could usefully be tested are the *Corpus of Contemporary American English* (COCA) and the *Corpus of Global Web-Based English* (GloWbE). The COCA was developed by Mark Davies and currently contains more than 450 million words, including 20 million words each year from 1990-2012 (as of August 22, 2014). It is a balanced corpus, being equally divided among five genres/registers: spoken, fiction, popular magazines, newspapers, and academic journals. Importantly, it is not static, as it is updated at least twice each year, which promises to keep it current, instead of being a 'snapshot' of English at a single point in time like the BNC. The COCA is thus an excellent corpus from which to derive vocabulary size information. It is now available to be fully downloaded onto one's personal computer at http://corpus.byu.edu/coca/, which makes the suggested replications eminently feasible.

The GloWbE is a brand-new corpus (also created by Mark Davies and released in April 2013) consisting of 1.9 billion words from 1.8 million web pages in 20 different English-speaking countries. Given the importance of the internet for global communication and information transfer, it would be very interesting to determine how much vocabulary

knowledge is necessary to comprehend this resource at the 95% and 98% levels of coverage, especially as it is so diverse and dynamic. It can be accessed at http://corpus2.byu.edu/glowbe/ and is also fully downloadable.

The *British National Corpus* (BNC) is still another corpus possibility. At 100 million words, it is smaller than either the COCA or GloWbE. It is also becoming dated, as it was compiled from a range of sources in the latter part of the 20th century. However, its 10-million word spoken component is relatively large for an unscripted spoken corpus, and this could be useful for exploring the requirements for spoken English. The BNC can be consulted at http://www.natcorp.ox.ac.uk/ or http://corpus.byu.edu/bnc/.

Another variable that could be explored is the word lists which are used to interrogate the target corpora. Nation (2006) used a word list based on the BNC, but this British-English based metric may be due for updating and revision. Nation recently took a step in this direction by updating his original BNC-based frequency lists using COCA frequency information. His goal was to further increase the generality of his original lists by reducing their British bias and making them more applicable to both British and American contexts. (See Nation, 2012a, available online, for details of the procedure.) The differences between the new and old lists are extensive and this has implications for the previously established coverage levels. Assuming the new combined BNC-COCA lists are a better indication of word frequency, then basically everything that has been done using the original BNC-based lists is ripe for replication using these new lists. Such replications may well change the established picture considerably. For instance, applying these more American lists could result in a downwards revision of the figure of 8,000 -- the word size needed to achieve the 98% coverage level according to Nation's (2006) investigation of materials that included the American novels *The Turn of the Screw* and *The Great Gatsby* and the American film script *Shrek.* In other words, in these replications, American-English

texts would no longer be analyzed using a British-English frequency scheme, and this might well reveal that the levels of coverage reported in Nation (2006) can be reached with smaller vocabulary sizes.

Another proposal pertaining to the counting unit used in investigations of coverage is to replace word families altogether, and use lemmas instead. Although lemma-based studies would not be comparable with the range of family-based studies the field is built upon, there would be a number of advantages. In particular, lemmas might be a suitable counting unit for research focusing on vocabulary pedagogy, as learners do not typically know all word family members (e.g. Schmitt ~~and~~ & Zimmerman~~,~~ 2002; see Schmitt~~,~~ 2010, Section 5.2.1, for a full discussion of pros and cons of different counting units). A lemmatized frequency list of the complete COCA has recently been made available (up to the first 100,000 words) by Mark Davies on his COCA website (http://www.wordfrequency.info/intro.asp). A comparison study of vocabulary size using one of Nation's word family wordlists and Davies' lemma wordlist would be interesting indeed, and would help in interpreting how generalizable Nation's word family figures are for pedagogical purposes. Since lemmatized lists are more easily made automatically via alphabetical grouping than family lists (which require manual work, e.g. to add *unhappy* to the *happy* family), this avenue of research could also identify important efficiencies in creating new lists from ever-evolving and dynamic corpora.[5]

An extension to the vocabulary size issue would be to determine whether Nation's vocabulary coverage and size figures as calculated for general English also pertain to more specialized domain-specific contexts. That is, can we generalize his vocabulary requirements for general English to more specific domains, e.g. within Academic English or Professional English? A small number of studies (e.g. Hsu~~,~~ 2011, Dang & Webb~~,~~ 2014) suggest that this is not particularly straightforward. The studies show that size figures for

the same levels of coverage differ depending on the degree of specificity. For example, Dang & Webb (2014) found that while 8,000 word families achieved 98% coverage of a general academic corpus, the subdiscipline requirements ranged from 5,000 (social sciences) to 13,000 (life and medical sciences). This implies that it is necessary to develop specific size requirements tailored to various domains and contexts. It is beyond the remit of this article on replication to give details about how to operationalize this extension of Nation's research. Instead, we direct readers' attention to Hsu (2014), who provides a useful model of how this type of research might be carried out. Focusing on engineering, she profiled a 4.57 million word corpus of English engineering textbooks. Using Nation's general BNC/COCA 25K lists, she found that knowledge of the first 2K plus proper nouns provided just 80.7% coverage of the corpus, and that 5,000 word families were needed to reach 95% coverage. Given that in Taiwan knowledge of 2,000 general service words is a high school graduation threshold, Hsu's aim was to develop a word list that would fill the 14.3% gap in coverage between the 2K level and 95% coverage. The resulting Engineering English Word List consists of just 729 word families, considerably fewer than the 3,000 general word families needed to achieve similar coverage. This study is a good example of how the principles of coverage and size discussed in our article can be used to deliver very focused and pedagogically useful vocabualry learning targets for particular domains and learning contexts.

2.3 Coverage, size, and learning (Cobb~~,~~ 2007)

Size and coverage can be used to calculate not just comprehension, but also how much vocabulary it is possible or probable to learn from a particular text. Comprehension and

learnability are interrelated, with the 95%-98% coverage figures commonly cited as "lexical thresholds" for both, although there is only one empirical finding that we know of to support the learning aspect. Swanborn ~~and~~ & De Glopper (1999) undertook a meta-analysis of 20 studies of word learning in L1 reading. They determined that known-unknown word ratios were a significant predictor of successful inferencing, and located some evidence for a threshold at one unknown word in 37 known (or, when 97% of the words are known). This coverage figure for successful inferencing is remarkably close to the 98% figure identified for L2 comprehension. However, beyond Swanborn ~~and~~ & De Glopper, the coverage-inferencing-learning link is largely a common sense intuition at this point, and would be a fruitful area for research. But what seems sure is that learners with a typical vocabulary size of 2,000 word families will not comprehend or learn much new vocabulary from a text with more than 10% of its vocabulary beyond the most frequent 3,000 word families in English (3K), as is typically the case with all but simplified texts. Nor will they consolidate any correct inferences they do manage to make with texts that recycle these words only once or twice, again as is typically the case.

Cobb's study looked at this learnability issue and started from the question, 'Can an adequate L2 reading lexicon be built from reading alone?' His methodology made the following choices:

- He used Nation's (2006) BNC-based frequency lists as a source of learning objectives, in this case the third 1,000 word families
- He hypothesized as the participant of his study a typical academic ESL learner with a vocabulary size of just over 2,000 words (2,112 word families, SD=1,036, is a rough international average for academic ESL learners, according to Laufer's

(2000) census of EAP (English for Academic Purposes) and ESP (English for Spcific Purposes) instructors in eight countries).

- He assumed a year of study as the learning period, one year being the typical allowance in ESP-EAP situations, or two in rare situations
- He assumed a maximum total yearly reading diet of either the Press, Academic, or Fiction divisions of the Brown corpus (179,000; 163,000; and 175,000 word tokens respectively, any of which equals about six stories the size of *Alice in Wonderland* or about 25 academic studies.The nature of the sub-corpora was intended to represent roughly the types of texts academic learners might be assigned to read, and the amount to be a generous estimate of what such students actually would read (remembering that with lexicons of 2500 words, such texts would be presenting at least one unknown word in ten and hence be rather arduous to get through).

The specific research question, then, was how many of the most frequent 3,000 word families in English (3K) are present in each of these collections, and in what coverage proportions, from the perspective of a learner with knowledge of just over 2,000 words. Random samples from each of these first, second, and third 1,000 levels of the BNC lists were matched against the contents of the three hypothesized reading diets. The finding was that while the first and second 1,000 word families are well represented in any of the diets, the third 1,000 families thin out rather dramatically in all of them, with only about half appearing even six times. (Eight to ten times seems to be the minimum figure for reliable incidental learning indicated by the research, e.g. Horst, Cobb & Meara, 1998). In other words, not much progress with the third 1,000 words could be expected from this presumably substantial exposure to natural (ungraded) text, at least not in the year or sometimes two that are normally available. By implication, then, pedagogies other than

reading alone are needed to assure adequate progress toward the coverage objectives discussed in earlier parts of this paper.

This somewhat pessimistic finding about vocabulary growth from reading is controversial to say the least. First, it goes against the view, once almost universally held and still common amongst practitioners, that contextual inference from self-selected input is sufficient for all levels of vocabulary development (e.g., Krashen, 1985). Second, it is a type of study that has not been undertaken before. Nation (2014, p. 1), called it a "notable exception" to a general lack of corpus-based studies of the feasibility of learning large amounts of foreign language vocabulary through reading, and, as already mentioned, it is novel in attempting to extend coverage analysis from comprehension to acquisition. Thus, for reasons of both controversy and originality, this study's findings will need substantial further investigation before they are accepted, ideally in the form of replication studies rather than just commentary and discussion. The replication should ideally be of two types, one that varies the data of Cobb (2007) and another that varies the assumptions about how or how much learners can read. Varying the data might involve, for example, pitching larger samples of second- and third-thousand words or other levels of words against different and possibly more representative corpora; varying the assumptions might involve basing calculations on different ideas about the amount of reading the hypothesized learners are able to perform in a year, or even empirical evidence, which would amount to replacing hypothetical with real learners. In fact, both types of replications of Cobb (2007) have already been undertaken, and these will now be summarized and any remaining questions identified.

On the corpus side, Nation in a series of conference presentations (2012b), a working paper entitled "How much input do you need to learn the most frequent 9,000 words?", and a forthcoming paper with the same title (2014), vastly extended the scale of

Cobb's (2007) corpus analysis, pitching each of the first nine k-levels against a 3-million token corpus of novels and other types of corpora, discovering in some detail the number of each level of words that would be encountered per unit of time assuming different reading rates. For example, if 300,000 words of the novels corpus were read, then 830 of the third 1,000 most frequent words would be met an average of 12.6 times (and so on up to just under 3 million words for most of the ninth 1,000), which is roughly in line with Cobb's proposal that with 167,000 words an insufficient number of third 1,000 words would be met for ~~stsematic~~ systematic acquisition. Nation, however, went on to propose that 300,000 words of reading is, in fact, possible, provided that (1) that reading goals were set higher than they are now (although "there is no published research to support [these proposed reading figures] for learners of English as a foreign language", p. 7); and (2) the texts involved "were at the right level for [the learners] so that the target words would make up around 2% or less of the running words in the text" (p. 7) with unknown words therefore met in a ratio of about one unknown in 50 known. Natural text, however, in most cases will *not* provide unknown words in such a friendly ratio, so the point is moot. For a learner who knows 2,000 word families, even a novel such as ~~one in Nation's corpus~~ Lady Chatterly will comprise almost 7% post-fourth 1,000 items, such that unknown words will be met in a ratio of at least one unknown in 15 (roughly one per two lines of text). Further, many types of texts are even more lexically sophisticated than the novels on which these figures are based, and still further come in less inference-friendly formats than the chronological flow of everyday events which typifies novels. To summarize, then, the corpus part of Nation's replication confirms and extends Cobb's initial and in retrospect pilot-level work, but the proposal part simply highlights the need for empirical work on how much of what type of texts learners can actually read

This second type of replication, on the learner side, is the topic of McQuillan & Krashen's (2008) direct response to Cobb (2007). The form of the replication began with a review of existing literature on real rather than hypothetical L2 readers and their reading rates. These researchers questioned whether Cobb's proposed reading diet was actually particularly large for typical L2 learners, citing 11 reading-rate studies showing that real learners with lexicons of about 2,000 words can read a lot more than Cobb's 179,000 words of unsimplified text over a year, indeed rather more like 517,000 words. This would then mean that such learners would meet most of the third 1,000 target words enough times for learning and consolidation to occur.

Yet on actual inspection of McQuillan & Krashen's sources, (Cobb~~,~~ 2008), the posited larger amount of reading turns out to be the reading of simplified materials, not of academic or otherwise authentic texts. Thus, while more can undoubtedly be read using simplified materials, this is no guarantee that the particular target lexis (the third 1,000 word families, or beyond) is actually present in such materials ~~(Cobb, 2008)~~. Most existing simplified texts focus on the first 1,000-2,000 word families, although other targets would be possible (see below). In other words, reading rates of these learners for authentic texts are presently unknown, and a true empirical replication of the Cobb (2007) study with valid information about reading diet remains to be done.

Both the counting-up and the empirical types of replications are worth doing. Indeed both Nation and McQuillan & Krashen are the two parts of the single replication that is needed: a corpus-based work-out of the approximate learning opportunities, across the mid-frequency zones, in a range of relevant text types, that was based on a validated assessment of how much and what type of text learners at different levels are actually able to read. It is quite likely that the two parts will be engaged separately and assembled subsequently, although they could be done together, perhaps in a new research paradigm

that might be called corporo-empirical research. It is important, however, that this work ~~- although complex -~~ should be performed. The question of L2 vocabulary growth, particularly with regard to the demands of advanced level reading, has been unresolved for decades, but has now come in range – is answerable in principle through research-informed corpus analysis.

Within these two main lines of replication, a number of variants ~~are imaginable and~~ could be usefully incoporated within an "approximate" approach that might modify or even reverse the findings of Cobb (2007). One possibility would be to vary the type of texts in the corpus. For example, a study might look at the learning opportunities in not just academic or other authentic texts but also in pedagogically modified texts. Indeed, Nation, in response to some of the issues raised above, has recently begun producing a complete set of graded readers that specifically target "mid-frequency" vocabulary (that of the third to eighth 1,000 levels, as defined by Schmitt & Schmitt~~,~~ 2012), such that significant numbers of word families in target K-level zones are met frequently and mainly in environments of 98% known words. This work is described in Nation & Anthony (2012), and in an undated information document by Nation ("-About mid-frequency readers," http://www.victoria.ac.nz/lals/about/staff/paul-nation). The first 13 of a projected 50 modified texts (nine fiction and four non-fiction), each ~~written down~~simplified to a fourth, sixth, and eighth 1,000 word families target level, have been completed and are available on ~~—~~Nation's website. When completed, these 50 texts will form a corpus that will be eminently suitable for a combined corpus and empirical replication of Cobb (2007) or indeed of Nation (2014). ~~How many word families are available for learning at different frequency levels, by learners at different vocabulary levels? How many words can learners read? How many words are learned?~~

Another variation on the text/corpus side that could form the basis of a useful replication would be to vary both text genre and degree of prior familiarity with the topic area. Neither Cobb's choice of Press, Academic, or Fiction sections of ~~the~~ Brown ~~corpus~~, nor Nation's choice of out-of-copyright novels, even with lexical redesign, is typical reading for today's academic ESL learner. A corpus of what such learners are in fact reading ~~might be a useful starting place to replicate~~ could provide a very interesting replication of Cobb (2007). A researcher might well find that academic ESL learners, reading in their domains, ~~and~~ where they are building an accumulating knowledge base revolving around a limited number of themes (doing narrow reading), can indeed read larger amounts than those proposed by Cobb and can indeed build their lexicons substantially through reading.

~~A number of useful replications of Cobb (2007) with v~~Variations on the learner/empirical side could be also undertaken, separately or in conjunction with variations on the corpus side. A particularly important element of the original study that could be usefully varied in an approximate replication is reading conditions. The assumption of convenience in Cobb (2007), as in all coverage studies that we know of including those discussed in earlier parts of this paper, is that the reading occurs in unassisted conditions. But this is no longer how very many people actually read, particularly young people and students, given the click-on definitions, text-to-speech renditions, and Google searches beckoning within the Web pages and PDF files on their computers and iPads. It is almost certain that more difficult texts can be ~~read~~ comprehended and more vocabulary learned from resource integrated texts than with self-contained texts, but how much more?~~.~~ ~~-~~Changing the condition from unassisted to resource-assisted reading, perhaps in potential interaction with various kinds or degrees of training, one might well find that learners with emergent lexicons of about 2,000 word families can indeed read texts like *Lady Chatterley* ~~can be read in volume and~~ with

comprehension, enjoyment, and substantial vocabulary growth ~~on the part of learners with vocabulary sizes of 2,000 words~~. Indeed, in the information document accompanying these designed-for-learning readers, Nation (n.d.) proposes precisely this method of engaging with his learning-enhanced texts ~~():~~ .

~~It is near certain that the size and coverage requirements will be lower under these conditions than for unassisted reading owing to the increased likelihood of accessing word meanings on the fly.~~ The interesting questions then become, ~~ones like:~~ 'How many words does one need to know to read texts connected to online resources?' and 'How many words can one learn from reading texts connected to online resources?' Cobb (2007) proposed that 170,000 words a year was a lot of reading for learners with 2,000 word families and that inferences of new word meanings would be hard going; but with resources, Nation's 300,000 or McQuillan & Krashen's 517,000 might well come in range, and easy look-ups make inferences straightforwardly confirmable. In other words, through these replications Cobb's gloomy prognosis might be shown to be an artefact of a now largely defunct unassisted reading paradigm.

3. Conclusion

Knowing how much vocabulary is necessary to be functional in a language is crucial for setting vocabulary learning goals and designing syllabuses. Setting vocabulary size goals in which the ELT community can be confident requires replication of both lexical coverage

and vocabulary size studies, with studies into learning potential a useful adjunct. We can probably use the 95% and 98% coverage figures for written language, but for spoken discourse, there is a clear need to build upon the initial findings of Van Zeeland ~~and~~ & Schmitt (2012). Nation (2006) provides clear learning targets for vocabulary size, but given the very considerable teaching and learning effort these substantial figures entail (6,000-7,000 word families for spoken discourse and 8,000-9,000 for written discourse), it is important to determine if these figures still obtain for other corpora, or if perhaps additional research will point to somewhat lower requirements (if we are lucky). Finally, we need to move beyond a merely corpus-driven discussion of the lexical needs for using language, and start looking into what learners can actually do with various vocabulary sizes, and how this affects their further learning, as Cobb (2007) suggests. All of these replications should lead to a firmer establishment of vocabulary size requirements necessary to inform language pedagogy and assessment.

Notes

1. Michel, et al. (2011) found 1,022,000 individual word forms in a computer analysis of a corpus of 5,195,769 digitised books.

2. There are a number of units with which to count vocabulary, with the most common being individual word form, lemma, and word family. These counting units have often been ill-defined or conflated in the literature, and so for convenience, we will use 'word' as our general cover term in this paper. In cases where a more specific counting unit is possible/appropriate (e.g. word family), we will use that term. Another complexity

concerns formulaic language. A great deal of language is made up of multiword units, but this formulaic language is seldom included in vocabulary counts. How to deal with formulaic language in counting, teaching, and testing vocabulary remains one of the most pressing issues in the field of vocabulary studies. However, as the measurement of the size and knowledge of formulaic language is still in its infancy, it will not be focused on in this paper dealing with replication. (See Schmitt, 2010 for details on all of these issues.)

3. One weakness of coverage figures is that they do not distinguish the importance of the individual words in a text. If unknown words are crucial to the understanding of a text (and this is likely for at least some), then high coverage figures will not by themselves ensure adequate comprehension. Research on the effect of particular words on comprehension has not yet been done, and is outside the remit of this paper on replication. However, the issue is probably central for the coverage argument, and needs to be developed as a complementary research strand.

4. The relatively congruent size requirements for a variety of written and spoken text types is partially a reflection of Zipf's law, where the most frequent words will always make up the majority of a text of any type.

5. Although there is not space to elaborate here, we feel that careful selection of appropriate counting units is essential for any future study, replication or not. Counting units need to be chosen in a principled manner, without word families being the automatic default just because they have been used before. It is also worth noting that although compiling word family lists requires an enormous amount of work in the first instance, once developed, their re-use requires relatively little editing.

Acknowledgements

We wish to thank Mélodie Garnier, Benjamin Kremmel, Marijana Macis, Kholood Saigh, Michael Rodgers, Hilde van Zeeland, Laura Vilkaite, and Paul Nation for their insightful comments on earlier versions of this paper.

References

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. Language Learning & Technology 11(3), 38-63.

Cobb, T. (2008). What the reading rate research does not show: Response to McQuillan & Krashen. Language Learning & Technology 12(1), 109-114.

Dang, T.G.Y. & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes, 33*, 66-76.

Field, J. (2004). An insight into listeners' problems: Too much bottom-up or too much top-down? *System, 32*, 363–377.

Goulden, R., Nation, P. & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics 11*, 341-363.

Horst, M., Cobb T., and Meara, P. (1998). Beyond A Clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*, 207-223.

Hsu, W. (2011). The vocabulary thresholds of business textbooks and business research articles for EFL learners. *English for Specific Purposes, 30*, 247-257.

Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes, 33*, 54-65.

Hu, M. & Nation, I.S.P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 23*, 403-430.

Krashen, S. (1985). The Input Hypothesis: Issues and Implications. London: Longman.

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

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

Lund, R.J. (1991). A comparison of second language listening and reading comprehension. *Modern Language Journal 75*, 196–204.

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

Michel, J-B., Yuan, K.S., Aiden, A.P., Veres, A., Gray, M.K., The Google Books Team, Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwan, J., Pinker, S., Nowak, M.A., Aiden, E.L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–82.

Nation, P. (n.d.). About the mid-frequency readers. Available at http://www.victoria.ac.nz/lals/about/staff/paul-nation/

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

Nation, P. (2012a). The BNC/COCA word family lists. Available at http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Information-on-the-BNC_COCA-word-family-lists.pdf.

Nation, P. (2012b). What does every ESOL teacher need to know? *CLESOL plenary for Language Teaching Community Languages and ESOL.* Palmerston North, NZ.

Nation, P. (2014). How much input do you need to learn the most frequent 9,000 words? Reading in a Foreign Language 26(2), 1-16.

Comment [T1]: TOM SAYS: A WANT TO KEEP THIS REFERENCE ASWELL AS THE 2014 BECAUSE IT IS PART OF THE 'TIME LINE' OF EXISTING RESPONSES TO THE 2007 PAPER. IF THE POINTS IT RAISES ARE CITED ONLY AS 2014, WE REDUCE CLARITY AND STORY LINE

Nation, P. & Anthony, L. (2012). Mid-Frequency Readers. Journal of Extensive Reading 1(1), 5-16.

Park, G. (2004). Comparison of L2 listening and reading comprehension by university students learning English in Korea. *Foreign Language Annals 37*, 448–458.

Porte, G. (2012). Introduction. In G. Porte (ed.), *Replication research in Applied Linguistics*. Cambridge: Cambridge University Press, 1-18.

Rubin, J. (1994). A review of second language listening comprehension research. *Modern Language Journal, 78,* 199–221.

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. Basingstoke: Palgrave Macmillan.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal 95*, 26-43.

Schmitt, N., Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*. doi:10.1017/S0261444812000018.

Schmitt, N. and Zimmerman, C.B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly, 36,* 145-171.

Swanborn, M. & de Glopper,K. (1999). Incidental word learning while reading: A meta-Analysis. Review of Educational Research 69 (3), 261-285.

Vandergrift, L. (2011). L2 listening: Presage, process, product and pedagogy. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning. Vol. II*. Routledge, pp. 455–471.

van Zeeland, H. & Schmitt, N. (2012). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*. doi:10.1093/applin/ams074.

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

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

Wellington Corpus of Spoken English. Information at <http://www.victoria.ac.nz/lals/resources/corpora-default/corpora-wsc>.