Research Investigating Lexical Coverage and Lexical Profiling: What We Know, What We Don't Know, and What Needs to be Examined

Stuart Webb
University of Western Ontario
Canada

Abstract

Studies of lexical coverage are valuable because they reveal the importance of vocabulary knowledge to comprehension. Lexical profiling research is also extremely useful because it indicates the vocabulary knowledge necessary to understand different text types such as novels, newspapers, academic lectures, television programs, and movies. Moreover, lexical profiling research provides teachers and learners with concrete vocabulary learning targets that students can seek to achieve and evaluate their knowledge against. However, there are only three studies that have precisely investigated the effects of lexical coverage on reading comprehension (Hu & Nation, 2000; Laufer, 1989; Schmitt et al., 2011), two that have directly investigated its effects on listening comprehension (Bonk, 2000; Van Zeeland & Schmitt, 2013), and one that has done this for viewing comprehension (Durbahn et al., 2020). With few studies and few variables that may affect comprehension examined, discussions of the generalizability of lexical coverage findings are likely overstated. The aim of this article is to clarify earlier research findings and highlight areas where further research is needed.

Keywords: vocabulary, lexical coverage, lexical profiling, reading comprehension, listening comprehension, learner variables, text variables

Lexical coverage refers to the degree to which words in input are known by readers and listeners. For example, if a text consists of 100 running words and five of those words are unknown, lexical coverage of that text would be 95%. Studies of lexical coverage have primarily investigated its relationship with comprehension with research indicating that as lexical coverage increases above 90%, comprehension is also likely to improve (Bonk, 2000; Hu & Nation, 2000; Laufer, 1989; Schmitt et al., 2011; Van Zeeland & Schmitt, 2013).

The lexical coverage figures that are frequently cited as indicative of adequate comprehension are 95% and 98% (Laufer & Ravenhorst-Kalovski, 2010). These lexical coverage figures might initially appear higher than expected; I have found that teacher training candidates often predict that around 80% lexical coverage may be sufficient for reading comprehension. Therefore, it is useful to consider the amount of effort that is required to understand written text to make the figures more transparent. For example, there are usually around 14 words in a line of typed text and 400 running words on a page. At 95% lexical coverage, one out of every 20 words is unknown which means that there are about two unknown words in every three lines on a page, and about 20 unknown words per page. At 98% lexical coverage, one out of every 50 words is unknown which means that there is about

one unknown word every 3.5 lines and about eight unknown words per page. Considering the amount of effort that would go into understanding 20 pages of written text and the 400 encounters with unfamiliar words at 95% coverage, lexical coverage figures of at least 95% for reading comprehension make sense.

Studies of lexical profiling investigate the vocabulary knowledge necessary to reach the 95% and 98% lexical coverage figures associated with comprehension. Lexical profiling research investigates how many words are needed to understand different types of discourse (e.g., Dang & Webb, 2014; Nation, 2006; Webb & Rodgers, 2009a, 2009b). Nation's (2006) seminal study of lexical profiling indicated that 6,000 to 7,000 word families were necessary to reach 98% lexical coverage of spoken text, and 8,000 to 9,000 word families were needed to reach 98% lexical coverage of written text. There has been variation in these vocabulary knowledge estimates in studies since then (e.g., Webb & Macalister, 2013), particularly for estimates of the vocabulary knowledge necessary to understand spoken discourse (e.g., Dang & Webb, 2014; Webb & Rodgers, 2009a, 2009b) which tend to use the 95% lexical coverage figure rather than 98%.

Discussions of lexical coverage and lexical profiling have become increasingly common (e.g., Laufer, 2013, 2020; Nurmukhamedov & Webb, 2019; Schmitt et al., 2017). Studies of lexical coverage have generated interest because they reveal just how important vocabulary knowledge is for comprehension. Studies of lexical profiling are valuable because they provide vocabulary learning targets that may indicate when a discourse type may be understood, might be used to motivate further vocabulary learning, and reveal the importance of lexical development in second language (L2) learning. However, because there are few studies of lexical coverage, researchers should be cautious when interpreting earlier findings. Recent discussions of the topic tend to overgeneralize earlier findings, making it difficult to evaluate the validity of their claims (e.g., Brown et al., 2020; McLean, 2021; Stoeckel et al. 2020). The aim of this article is to clarify earlier research findings and highlight areas where further research is needed.

What do studies of lexical coverage reveal?

Three studies (Hu & Nation, 2000; Laufer, 1989; Schmitt et al., 2011) have precisely investigated the relationship between lexical coverage and reading comprehension. In the first study to investigate lexical coverage, Laufer (1989) found that there were a greater number of English as a foreign language (EFL) learners who achieved reading comprehension scores of 55% or higher at 95% lexical coverage than those who did not. Hu and Nation (2000) found that no EFL learners had adequate comprehension (12 or higher out of 14 on a multiple-choice test and 70 or higher out of 124 on a cued written recall test) at 80% coverage, few EFL learners achieved adequate comprehension at 90% coverage, and the majority of participants did not achieve adequate comprehension at 95% coverage. At 100% coverage, 15 and 10 of the 17 participants achieved adequate comprehension on multiple-choice and cued meaning recall comprehension tests, respectively. In the most comprehensive study of the topic, Schmitt et al. (2011) found that there was a linear relationship between lexical coverage and reading comprehension with reading comprehension scores tending to improve as lexical coverage increased above 90%. Based on the initial findings, Laufer (1989) suggested that 95% lexical coverage may be sufficient for reading comprehension, while Hu and Nation (2000) and Schmitt et al. (2011) suggested 98% lexical coverage was most appropriate. Laufer and Ravenhorst-Kalovski (2010) argued that the amount of lexical

coverage required should be a function of the degree of comprehension that is expected. They suggest that 95% lexical coverage may be sufficient for 'minimally acceptable' understanding, but 98% coverage may provide a more precise understanding.

Two studies (Bonk, 2000; Van Zeeland & Schmitt, 2013) have directly investigated the relationship between lexical coverage and listening comprehension. Bonk (2000) found a great deal of variation in listening comprehension scores at different levels of coverage. Comprehension tended to be best with lexical coverage above 90%. However, some participants also achieved good listening comprehension between 80% to 89%. Van Zeeland and Schmitt (2013) found that the majority of L2 learners had adequate comprehension of a listening passage at 90% (7.35/10) and 95% (7.65/10) lexical coverage (means at 98% and 100% lexical coverage were 8.22 and 9.62, respectively). However, because there was less variation in comprehension scores at 95% coverage than at 90%, they recommended that 95% coverage was most appropriate for listening comprehension. There has also been one study examining the relationship between vocabulary knowledge and viewing comprehension. Durbahn et al. (2020) found that EFL learners' comprehension of a 30-minute television documentary increased from 62% at 87% lexical coverage to 92% at 99% lexical coverage. However, comprehension scores did not always increase between 87% and 99% lexical coverage, perhaps due to a relatively small number of participants.

There are several other studies that have indirectly examined lexical coverage and comprehension (Laufer, 1992; Laufer & Ravenhorst-Kalovski, 2010; Noreillie et al., 2018; Stæhr, 2009). Laufer (1992) and Laufer and Ravenhorst-Kalovski (2010) investigated the relationship between vocabulary knowledge and reading comprehension. Laufer (1992) showed that higher scores on tests of vocabulary knowledge (Nation's (1983) version of the Vocabulary Levels Test and Meara and Jones' (1989) Eurocentres Test) were linked with improved reading comprehension. Laufer and Ravenhorst-Kalovski (2010) did the same but included an estimation of the lexical coverage for participants with approximate vocabulary sizes. Stæhr (2009) and Noreillie et al. (2018) looked at the relationship between vocabulary knowledge and listening comprehension and estimated lexical coverage by comparing lexical profiles of the listening material to scores on tests of vocabulary knowledge. The results of the three studies indicated that as participants' vocabulary knowledge and estimated lexical coverage increased, so did their mean comprehension scores.

Together, Laufer and Ravenhorst-Kalovski (2010), Stæhr (2009), and Noreillie et al. (2018), as well as other studies that have explicitly investigated the relationship between vocabulary knowledge and comprehension (e.g., Qian, 1999, 2002; Stæhr, 2008, 2009), support the notion that vocabulary knowledge may be the key factor in whether or not written text is understood. However, although these studies show that learners with greater vocabulary knowledge tend to have better reading comprehension than those with less vocabulary knowledge, there are two reasons why we should be cautious about attributing improved comprehension to increased lexical coverage. First, lexical coverage of the input is based on the degree to which participants demonstrated knowledge of different word frequency levels rather than on their knowledge of the vocabulary encountered in the texts. This means that participants' actual lexical coverage of the input is at best a very rough estimation. Second, although the studies revealed that mean comprehension scores significantly increased in relation to increased vocabulary knowledge, none of the studies investigated the degree to which individual participants with different amounts of vocabulary knowledge and estimated lexical coverage achieved a particular level of comprehension.

Together, the studies of lexical coverage tend to indicate that as knowledge of the vocabulary within text increases above 90%, so does comprehension (Bonk, 2000; Hu & Nation, 2000; Laufer, 1989; Schmitt et al., 2011; Van Zeeland & Schmitt, 2013). However, the lexical coverage figures associated with comprehension may vary between written, spoken, and audiovisual input. Lexical coverage of 90%, 95%, and 98% may be sufficient for audiovisual, listening, and reading materials to be understood, respectively. However, the degree to which text will be understood is likely to fluctuate to some degree among learners above and below these figures. Schmitt et al.'s (2011) study clearly shows that some learners could achieve a high level of reading comprehension at lexical coverage as low as 90%. It is also extremely important to note that studies consistently showed that 100% coverage does not ensure that spoken and written texts are understood. The most positive findings at 100% lexical coverage were for listening comprehension. Van Zeeland and Schmitt (2013) found that 29/40 participants had perfect listening comprehension test scores and no participants scored lower than 7/10 at 100% lexical coverage. In contrast, Hu and Nation (2000) found that at 100% coverage, only two out of 17 participants had perfect scores on their reading comprehension test, while Schmitt et al. (2011) found that at 100% coverage, the mean scores on two texts were 19/30 and 23/30. Schmitt et al. also found that even with 100% coverage, participants scored as low as 7/30 and 9/30 on comprehension tests for their two reading passages. Thus, while reaching target lexical coverage figures might indicate that text might be comprehensible, it does not ensure comprehension. Many other factors also influence the comprehensibility of spoken (Boyle, 1984), written (Grabe, 2009), and audiovisual input (Pujadas & Muñoz, 2020), and studies of lexical coverage clearly reveal that the degree to which something is understood at 90%, 95%, 98%, and 100% coverage will vary among learners.

Which areas of research on lexical coverage deserve further investigation?

With relatively few studies of lexical coverage, little is known about how different factors moderate the relationship between lexical coverage and comprehension. Schmitt et al. (2011) found that greater background knowledge tended to lead to increased reading comprehension as lexical coverage increased from 94% to 100%. However, the degree to which other learner factors moderate comprehension remains to be examined. Studies of lexical coverage have included relatively advanced participants (Hu & Nation, 2000; Schmitt et al., 2011; Van Zeeland & Schmitt, 2013), and it is not clear whether their findings apply to lower-level learners. It may be that more advanced learners have greater syntactic and metacognitive knowledge, and depth of vocabulary knowledge, which also impacts their ability to understand text. Examining the effects of lexical coverage on comprehension with learners of varying proficiencies and vocabulary levels would be a useful follow-up to the earlier studies.

The extent to which text factors moderate the effects of lexical coverage on comprehension also deserves further investigation. Studies that have precisely measured lexical coverage have consisted of one text that was academic in nature (Laufer, 1989), a relatively easy narrative text of 673 running words (Hu & Nation, 2000), two texts (582 and 757 running words) of similar difficulty that were academic in nature, but with topics that varied in topic familiarity (Schmitt et al., 2011), four listening passages of 84-86 running words (Bonk, 2000), four short stories of 470-485 running words (Van Zeeland & Schmitt, 2013), and a 30-minute documentary (4,626 running words: Durbahn et al., 2020). We should expect that factors, such as text length, text type (narrative vs expository), and text genre (written: short passages from course books, graded readers, news articles, novels, academic articles; spoken:

short listening passages from course books, radio programs, podcasts, monologues, dialogues; audiovisual: documentary, comedy, and drama television programs, movies, Youtube videos, Ted Talks), moderate the relationship between lexical coverage and comprehension to some extent. Based on the very small number of texts and text types that have been included in research, we should be cautious to generalize the findings of earlier studies to other text types. Further research examining the relationships among text factors, lexical coverage, and comprehension is clearly warranted.

It would also be useful to investigate the degree to which methodological variables moderate the relationship between lexical coverage and comprehension. For example, it would be interesting to examine how different approaches to determining lexical coverage influence findings. Research involved participants underlining the unknown words in passages (Laufer, 1989), replacing low frequency words with nonwords (Hu & Nation, 2000; Van Zeeland & Schmitt, 2013), and completing tests to measure participants' knowledge of the words in the materials (Bonk, 2000; Durbahn et al., 2000; Schmitt et al., 2011). There have been several different types of assessment used to determine known and unknown words encountered in the L2 input. Schmitt et al. (2011) used a checklist test. Bonk (2000) used a dictation test, and Durbahn et al. (2020) used a meaning recall test. Another methodological variable that should be considered is how repetition of the input affects findings. For example, participants in Van Zeeland and Schmitt's (2013) study listened to passages twice. Since research indicates that repeated reading (Gorsuch & Taguchi, 2008) and listening (Chang & Read, 2006) improve comprehension, it would be useful to determine the degree to which repeated reading, listening, and viewing moderate the relationship between coverage and comprehension.

Another area of research that is needed is examining how learners' reading behavior is affected by lexical coverage. Although research has shown that increased lexical coverage allows learners to better understand text, it would be useful to investigate how reading behavior and text processing changes at different levels of coverage, and how this relates to improved comprehension. It may be that at lower levels of coverage, attention is split between trying to understand the many unknown words in the text, as well as the text as a whole. Cognitive Load Theory (Sweller, 1988) would suggest that cognitive resources could be more effectively allocated to attending to the aspects of texts that are not understood at increased coverage levels. Examining the processing of text at different levels of coverage through global eye movement measures would paint a clearer picture of how learners' reading behavior is affected by lexical coverage. In addition, there would be great value in investigating the degree to which learners use additional resources, such as dictionaries at different levels of coverage to support comprehension. This would reveal the extent to which support for comprehension is necessary at different coverage levels. Laufer and Ravenhorst-Kalovski (2010) suggested that 95% coverage may be sufficient for reading comprehension with support, and at 98% coverage support may not be necessary. However, this remains to be explored in research. Examining the use of support in relation to lexical coverage levels would also shed light on the degree to which fluency is impacted by lexical coverage. As it stands, there does not appear to be any research looking at how lexical coverage impacts fluency. This is another area where research is needed because it has been argued that 98% lexical coverage is appropriate for extensive reading and 99-100% coverage is suitable for speed reading (e.g., Nation, 2007; 2013).

A final area where further research is needed is investigating the degree to which vocabulary can be learned at different levels of coverage (Webb, 2016). Two studies have looked at the degree to which words could be inferred during reading. Liu and Nation (1985) found that a

greater percentage of unknown words were inferred at 96% coverage than at 90%. However, because the texts were at two different lengths, it is not clear whether the results were attributable to lexical coverage or other text factors. In a small-scale exploratory study, Laufer (2020) also compared lexical inferencing at different levels of coverage and found that a significantly greater percentage of unknown words could be successfully inferred at 95% and 98% coverage than at 90% coverage. Participants inferred 52% of unknown words at 90% coverage, and 80% and 82% of unknown words at 95% and 98% lexical coverage, respectively. The degree to which words can be incidentally learned at different levels is yet to be examined and would be another useful follow-up to these studies.

What do lexical profiling studies reveal?

The development of Nation's (2006) British National Corpus word lists, Nation's (2012) British National Corpus/Corpus of Contemporary American English word lists, and lexical profiling software, such as the RANGE program (Heatley, et al., 2002), AntWordProfiler (Anthony, 2014), and Vocabprofile (Cobb, 2021), made it possible to determine the number of word families necessary to reach the lexical coverage points associated with reading, listening, and viewing comprehension. Subsequently, there have been many studies investigating the number of word families (in sets of 1,000) needed to understand different types of discourse (for a brief description and list of these studies, see Nurmukhamedov & Webb, 2019).

Table 1

Lexical Coverage of High Frequency and Mid Frequency Word Families

Source	Text type	1-1,000	1,001-2,000	2,001-3,000	3,001-9,000
Webb & Macalister (2013)	Graded readers	91.06	5.67	1.79	1.18
Webb & Macalister (2013)	Written text	82.43	7.33	3.15	5.07
Webb & Paribakht (2015)	CanTEST reading passages	74.69	11.33	4.30	6.37
Webb & Rodgers (2009a)	Television	89.10	4.42	1.93	3.23
Tegge (2019)	Songs	89.70	4.22	1.95	2.96
Dang & Webb (2014)	Academic speech	87.54	5.40	1.76	3.55
Webb & Paribakht (2015)	CanTEST listening passages	81.09	8.74	3.13	4.89

Lexical profiling research has demonstrated the relative value of the most frequent word families across different discourse types. Table 1 shows the lexical coverage of high frequency and mid frequency vocabulary for several discourse types. The table shows that the most frequent 1,000 word families account for most of the words that are encountered in English, about 75% to 91% of spoken and written text. The coverage of the next two sets of 1,000 word families is far less. The most frequent 1,001-2,000 word families account for about 4% to 11%, and the most frequent 2,001-3,000 word families cover about 2% to 4% of spoken and written text. In contrast, lexical coverage of the next 6,000 word families as a whole accounts for only about 1%- 6% of spoken and written text. These studies highlight the importance of learning vocabulary in relation to word frequency levels and the value of learning the most frequent words to facilitate comprehension.

Written text

Lexical profiling studies also reveal large differences between the lexical demands of reading materials. This is useful because it indicates which types of materials may be most easily understood, as well as vocabulary learning targets that may facilitate comprehension of discourse types. For example, Nation (2006) found that the most frequent 8,000 and 9,000 word families plus proper nouns provided 98% coverage of newspapers and novels, respectively. In addition, Webb and Macalister (2013) found that the most frequent 3,000 word families plus proper nouns was sufficient to reach 98% coverage of a corpus of graded readers, whereas 10,000 word families plus proper nouns were needed to reach 98% lexical coverage of both the Wellington Written Corpus, as well as a corpus of first language (L1) materials written for children.

Two studies examined the vocabulary profiles of reading passages in established English proficiency tests. This is useful because it indicates the vocabulary knowledge that may be required to do well in reading comprehension test items. Kaneko (2014) found that 12,000 to 13,000 word families plus proper nouns were needed to reach 98% lexical coverage of passages in the reading comprehension section of the TOEFL iBT test. In addition, Webb and Paribakht (2015) found that 14,000 word families plus proper nouns were needed to reach 98% coverage of reading passages in the CanTEST English Proficiency test.

Together, lexical profiling studies of written text reveal that (a) materials, such as graded readers that are written for L2 learners, are more appropriate for use as meaning-focused written input than materials written for L1 speakers, (b) knowledge of at least the most frequent 8,000 word families plus proper nouns might be an appropriate vocabulary learning goal for unassisted comprehension of materials written for L1 speakers, (c) lexical development should play a key role in language learning programs that aim to boost reading comprehension, and (d) developing the vocabulary knowledge necessary for unassisted reading of materials written for L1 speakers of English is likely beyond the scope of most current EFL learning programs. Although it may be possible for a small number of EFL students to learn the most frequent 8,000 word families by the end of their university studies, the vast majority of EFL students may not come close to achieving this goal. Research indicates that after as many as nine years of formal English language instruction, many EFL learners are still struggling to learn the most frequent 1,000 and 2,000 word families (Webb & Nation, 2017). This suggests a need to rethink and revise EFL learning programs.

Spoken discourse

Studies investigating the vocabulary in spoken and audiovisual materials have consistently shown that knowledge of the most frequent 3,000 to 4,000 word families plus proper nouns is sufficient to reach 95% lexical coverage. Nation (2006) found that 3,000 word families plus proper nouns provided 95% coverage of a corpus of talk-back radio and interviews and another corpus of conversation. Adolphs and Schmitt (2003) found that the most frequent 3,000 word families within a corpus of spoken English provided 95% coverage of that corpus. Tegge (2017) found that 3,000 word families plus proper nouns provided 95% coverage of a corpus of 408 pop songs. However, she also found that songs selected by teachers to aid learning needed only 2,000 word families plus proper nouns to reach 95% coverage. Dang and Webb (2014) found that 4,000 word families plus proper nouns provided 95% coverage of a corpus of academic spoken discourse. Two studies have examined the vocabulary profiles of listening passages in established English proficiency tests. Kaneko (2015) found that 3,000 word families plus proper nouns were needed to reach 95% coverage

of listening passages on the listening section of the TOEFL iBT test. Webb and Paribakht (2015) found that 4,000 word families plus proper nouns were needed to reach 95% coverage of listening passages in the CanTEST English Proficiency test.

It is useful to consider the lexical profiles of audiovisual input together with different forms of aural input because the language encountered in scripted television programs may also reflect the vocabulary used in conversation (Quaglio, 2009). A series of studies indicated that 3,000 word families plus proper nouns provides 95% coverage of television programs (Rodgers & Webb, 2011; Webb, 2010; Webb & Rodgers, 2009a) and movies (Webb & Rodgers, 2009b). Two studies investigated the lexical profiles of TED Talks. Both Coxhead and Walls (2012) and Nurmukhamedov (2017) found that 4,000 word families plus proper nouns provided 95% coverage of corpora of TED Talks. Taken together, lexical profiling of spoken and audiovisual input suggests that knowledge of the most frequent 3,000 word families plus proper nouns might be an appropriate vocabulary learning goal for unassisted listening and viewing comprehension because it will often be sufficient to reach 95% lexical coverage of listening materials.

Considered as a whole, lexical profiling of spoken and written input is important for two reasons. First, it indicates vocabulary learning goals for unassisted comprehension that are far smaller for spoken input (the most frequent 3,000 word families) than for written input (the most frequent 8,000 word families). This suggests that (a) there should be substantial focus in L2 learning programs to help students learn the spoken forms and meanings of the most frequent 3,000 word families and to support comprehension of spoken and audiovisual input, and (b) the use of extensive listening and viewing (together with extensive reading and language focused vocabulary learning) may be particularly useful in helping L2 learners to develop the vocabulary knowledge required for unassisted reading comprehension. A second reason why lexical profiling of spoken and written input is important is that it provides concrete vocabulary learning targets that students and teachers can work towards achieving. L2 learning is a long process in which improvement can be difficult for learners to see. Although these vocabulary learning goals cannot ensure that spoken or written discourse is understood (lexical profiles of individual texts vary and there are many factors besides vocabulary that affect comprehension), with support and training in the classroom, reaching these targets may soon lead to unassisted comprehension. Vocabulary Levels Test (Webb et al., 2017) scores can reveal to teachers and learners on an annual basis, where students are and how much more they need to learn to reach some of the vocabulary learning targets indicated by profiling studies. This may motivate further vocabulary learning.

Which aspects of lexical profiling research need clarification or deserve further investigation?

The vocabulary learning goals indicated by lexical profiling studies are useful. However, it is important to be aware that these goals provide an indication that a discourse type might be understood. They do not ensure that the discourse type will be understood. There would be great value in exploring the extent to which learners who have achieved the vocabulary learning goal for a discourse type can understand those materials to clarify the degree to which these lexical targets are meaningful. This would provide a more ecologically valid investigation of the degree to which lexical profiling and coverage might be useful for teaching and learning. There are also several other areas of lexical profiling research that deserve clarification or further study.

Lexical profiling studies should reveal the variation in lexical coverage among texts

The vocabulary learning targets determined through lexical profiling reflect the mean number of word families needed to reach a certain lexical coverage figure. However, it is important to be aware that each corpus is made up of a large number of texts, and there is likely to be a great deal of variation in the vocabulary of each text. For example, Webb and Macalister's (2013) analysis of children's literature indicated that knowledge of the most frequent 10,000 word families plus proper nouns and interjections was sufficient to reach 98% coverage. However, their analysis also looked at four sections of the corpus and showed that (a) the most frequent 9,000 word families plus proper nouns and interjections provided 98% coverage of one section, (b) the most frequent 10,000 word families plus proper nouns and interjections was sufficient to reach 98% for another section, and (c) the most frequent 11,000 word families were needed for the other two. This variation in the lexical profiles of texts within corpora is also reflected in several other studies (e.g., Dang & Webb, 2014; Webb & Paribakht, 2015; Webb & Rodgers, 2009a, 2009b). Thus, the learning target that is indicated by the analysis of a corpus will not always reflect the vocabulary knowledge needed to understand individual texts for that discourse type. In addition, the shorter the text, the greater the potential for variation in lexical coverage. For example, the coverage provided by 10 unknown words in texts that are 100, 1,000, and 10,000 words long is 10%, 1%, and .01%, respectively. This means that there is likely to be the most variation in lexical coverage of shorter texts such as news articles and short passages, because their lexical coverage figures can fluctuate greatly by the inclusion of a relatively small number of low frequency words. In contrast, less variation is likely to be found in longer texts, such as novels. This is likely the reason why there has been variation in the vocabulary learning targets derived from different corpora; a smaller number of longer non-technical texts is likely to lead to a lower vocabulary learning goal because there is less variation in topics (Sutarsyah et al., 1994), and in turn, the words in the corpus (e.g., novels in Nation, 2006) than a larger number of shorter texts (e.g., the Wellington Written Corpus in Webb & Macalister, 2013; reading comprehension test passages in Kaneko, 2014 and Webb & Paribakht, 2015).

One limitation of many lexical profiling studies is that the proportion of texts within a corpus that reflects the vocabulary learning target is not provided (e.g., Dang & Webb, 2014; Webb & Paribakht, 2015; Webb & Rodgers, 2009a, 2009b). It would be useful for future studies to include detail of the variation in the lexical profiles of the texts within the corpus to clarify the degree to which target vocabulary knowledge figures are likely to be accurate, too high or too low for the text type under investigation. This would help to make findings more transparent and clarify the potential for selecting texts that are at different vocabulary levels.

Lexical profiling studies should provide greater methodological transparency

Lexical profiling software, such as RANGE (Heatley et al., 2002) and AntWordProfiler (Anthony, 2014), classify the vocabulary in corpora according to the word lists (e.g, Nation's (2012) BNC/COCA lists) selected to use with the software. The lexical profiler classifies any words within the corpora that are not included in the word lists into a separate list that researchers must then examine to see if there are any words that were misclassified. Items in this list may include hyphenated words (e.g., year-old, sky-high), compound words (e.g., turnabout, timeslot), low-frequency words (e.g., subdural, sulfites), and proper nouns (e.g.,

Tylenol, Thurston) that were not found in any of the word lists. Researchers can then go through the process of making adjustments to the text or corpora that is analyzed (e.g., replacing hyphens with spaces so that year-old is reclassified as year and old) and the word lists (e.g., adding proper nouns to Nation's (2012) list of proper nouns) and then reanalyzing the corpus with the lexical profiler. The degree to which these adjustments are made affect the lexical profile. Therefore, it would be useful for there to be greater transparency about this process. Perhaps the best way to do this would be to make the lexical profiler output document (i.e., the classification of all words in the corpus according to the different word lists) available in an open open-access repository, such as the IRIS Digital Repository of Data Collection Materials (http://www.iris-database.org).

It is not clear whether matching vocabulary test scores to the lexical profiles of materials provides an accurate measurement of lexical coverage

It has been suggested that vocabulary size and levels test scores at different frequency levels can be matched against the lexical profiles of texts to determine their lexical coverage and in turn their suitability for use with students (Webb & Nation, 2008). In addition, several recent articles have been based on the assumption that vocabulary tests matched with the lexical profiles of materials indicate comprehensibility (e.g., Brown et al., 2020; McLean, 2021; Stoeckel et al., 2021). However, there is no research that has investigated the extent to which matching vocabulary size and levels test scores to lexical profiles provide an accurate measurement of lexical coverage, and there are several reasons why we should be cautious to assume that vocabulary size and levels test scores provide an accurate measurement of lexical coverage, and in turn comprehension.

First, it is important to note that tests of vocabulary size (e.g., Nation & Beglar, 2007) and levels (e.g., Webb et al., 2017) provide a very narrow measurement of vocabulary knowledge, which may not reflect the knowledge required to understand the vocabulary in any particular text. For example, size and levels tests measure knowledge of the form-meaning connections of individual words (e.g., run = move quickly). However, when these words combine with other words in a text, they may form unfamiliar collocations (e.g., run out of, run into, run the risk of), and this may have a negative impact on reading comprehension (Martinez & Murphey, 2011). Lexical profiling has not indicated the degree to which individual words and formulaic sequences in which they occur are present. This would be a useful area for further research.

Second, vocabulary levels and size tests evaluate the extent to which test takers can recognize the most frequent meanings of target items (e.g., run = move quickly). Although these meanings are the most likely to be encountered in input, to some extent secondary and peripheral meanings that may be unknown will also be encountered (e.g., run = manage; operate; cost). In contrast, lexical profiling software identifies words according to their individual forms rather than their different meanings or combined meanings with other words. Therefore, all primary, secondary, and peripheral meanings of a word within a corpus are counted as being the same. The degree to which polysemy affects comprehension would be a useful area for further research. For example, it would be useful to examine the degree to which the inclusion of secondary and peripheral meanings within text moderates the effects of lexical coverage on comprehension.

Third, scores on vocabulary size tests (e.g., Nation & Beglar, 2007) are not the same as the vocabulary learning goals derived through studies of lexical profiling. Lexical profiling

studies indicate a hypothetical vocabulary size that is made up of only the most frequent words. For example, Webb and Rodgers (2009a, 2009b) suggested that knowing the most frequent 3,000 word families plus proper nouns may be sufficient to understand television and movies. However, this learning target is different from a vocabulary size of 3,000 word families. The reason for this is that the vocabulary learning targets provided by lexical profiling studies consist of the highest frequency words which provide the greatest amount of coverage. Although words tend to be learned according to their frequencies (e.g., Schmitt et al., 2001), they are not learned in a rank order according to their frequency. Instead, mid- and low-frequency words will be learned along with high frequency words (Webb & Chang, 2012). Thus, a vocabulary size of 3,000 word families may be made up of a large proportion of the most frequent 3,000 word families, which will account for a great deal of coverage, and also a small proportion of words that are less frequent, account for very little coverage, and are less useful for comprehension. Thus, a vocabulary size of 3,000 word families will provide less lexical coverage than a vocabulary learning target of the most frequent 3,000 word families.

Researchers have provided mean comprehension scores according to vocabulary levels (Noreillie et al., 2018; Stæhr, 2009) and summed levels (Laufer & Ravenhorst-Kalovski, 2010) together with the lexical profiles of materials. It would be interesting for future research to examine the degree to which achieving mastery of vocabulary levels or summed levels in relation to the lexical profile of text indicates comprehension. Researchers have often discussed the importance of achieving mastery of vocabulary levels due to the lexical coverage that the higher frequency levels provide (e.g., Webb & Rodgers, 2009a, 2009b). However, there is yet to be research indicating that it is a better predictor of reading and listening comprehension than vocabulary size or summed levels.

Fourth, it may be that the lexical profiles of text are not a useful indicator of comprehensibility. Webb and Paribakht (2015) compared the lexical profiles of 38 reading comprehension passages and 37 listening comprehension passages at each of the 14 1,000-word levels in Nation's (2006) British National Corpus lists against the comprehension scores of 6,380 test takers for those passages. Because analyses of the lexical profiles of the passages revealed a great deal of variation in the vocabulary knowledge required to reach 95% and 98% lexical coverage, we might expect there to be relatively high correlation between their lexical profiles at different word frequency levels and comprehension of the passages; passages that included larger proportions of high frequency words might be better understood than those that included smaller proportions of high frequency vocabulary. Instead, there were either no correlations or small correlations between the two variables indicating that reading and listening comprehension of short test passages might be more largely affected by other variables. This led Webb and Paribakht (2015) to suggest that perhaps vocabulary knowledge might be a better predictor of comprehension than lexical coverage.

Fifth and most importantly, there is no research that has investigated the degree to which vocabulary test scores matched against the lexical profiles of text accurately measure lexical coverage. Vocabulary levels test scores provide an indication of student knowledge of different word frequency levels for the purpose of demonstrating the words that need to be learned. It seems unlikely that any test that estimates knowledge of 1,000 word frequency levels or vocabulary size will precisely indicate knowledge of words in a text. This is because tests that estimate knowledge of large numbers of words are likely to have some degree of error, and knowledge of a very small number of topic-related words in a text can account for

a significant amount of lexical coverage. This is not a limitation of the vocabulary tests because they were not designed for this purpose. However, it is reason to question the degree to which matching test scores to lexical profiles of materials may accurately indicate comprehensibility.

Whether there is value in creating more comprehensive vocabulary tests to try to provide more precise measurements of lexical coverage is not clear. Expecting test scores matched to lexical profiles of reading and listening materials to reveal comprehension might overestimate the degree to which vocabulary tests can measure lexical coverage. Moreover, it may also overestimate the degree to which vocabulary tests and lexical profiles can be used effectively by teachers in the language learning classroom. Vocabulary test scores may indicate which students are more and less likely to understand spoken and written text because those with more vocabulary knowledge will have more texts that reach target levels of lexical coverage. However, individual differences in vocabulary test scores and reading and listening comprehension skills among learners in each class, as well as large variation in the lexical profiles of text, make it difficult to apply lexical coverage figures to classes as a whole. In all likelihood, there will be varying degrees of text comprehensibility and the degree to which text is understood will be a function of learner variables (e.g., vocabulary knowledge, reading and listening skills, background knowledge) and text factors (e.g., lexical coverage, text length, text type, text genre, amount of formulaic language within the text). Brown et al. (2020), McLean (2021), and Stoeckel et al.'s (2021) discussions based on the assumption that vocabulary tests matched with the lexical profiles of a text indicate comprehension are at best misleading. Investigating the extent to which vocabulary levels and size test scores matched against the lexical profiles of different texts indicate comprehensibility would be a more pedagogically useful follow-up to the earlier studies of lexical coverage.

To what extent do learners with different vocabulary levels understand proper nouns and marginal words in spoken and written discourse?

The vocabulary learning targets identified by lexical profiling studies are typically described as the most frequent *n* word families plus proper nouns (e.g., Paul, Tokyo, Toyota) and marginal words (e.g., oh, ah, hmm, huh). Thus, when recommending a vocabulary learning target, there is often the assumption that if learners have reached that goal (at least the most frequent 3,000 word families), then they will also be able to understand the proper nouns that are encountered in the discourse type. Support for this assumption is based on the fact that proper nouns are likely easier to understand than content words because (a) they are usually indicated by capitalization in written input, (b) they are not typically expected to be known, and (c) their meanings are often revealed within the text (Nation & Webb, 2011). Although we might expect learners who have learned the most frequent 3,000 word families to understand proper nouns and marginal words, there is also reason to question whether proper nouns are understood (Brown, 2010), particularly in spoken discourse where they may be less transparent (Kobeleva, 2012). Thus, it would be useful to investigate the extent to which learners with different vocabulary levels (e.g., knowledge of the most frequent 1,000, 2,000, and 3,000 word families) can understand proper nouns and marginal words in spoken and written discourse.

Conclusion

Lexical coverage and profiling studies have generated a great deal of interest in recent years for good reason. They reveal the importance of lexical development for unassisted comprehension and provide useful vocabulary learning goals that learners can evaluate their knowledge against, and work to achieve. However, there are very few studies of lexical coverage, and there is a need to be cautious about the degree to which we generalize earlier findings to spoken, written, and audiovisual comprehension. Recent studies (e.g., Brown et al., 2020; McLean, 2021; Stoeckel et al., 2021) likely overestimate the degree to which we can generalize earlier findings to teaching and learning. With so few studies of lexical coverage, there are many potentially useful areas for further research. This article has tried to highlight several of the different topics that deserve investigation.

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425–438.

Anthony, L. (2014). AntWordProfiler (Version 1.4.1) [Computer Software]. Waseda University. Retrieved from http://www.laurenceanthony.net/

Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening*, *14*(1), 14–31.

Boyle, J. P. (1984). Factors affecting listening comprehension. *ELT Journal*, *38*(1), 34–38.

Brown, D. (2010). An improper assumption? The treatment of proper nouns in text coverage counts. *Reading in a Foreign Language*, *22*, 355–361.

Brown, D., Stoeckel, T., McLean, S., & Stewart, J. (2020). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, amaa061. https://doi.org/10.1093/applin/amaa061

Chang, A. C. S., & Read, J. (2006). The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly*, *40*(2), 375–397.

Cobb, T. *Compleat Web VP v.2.5* [Computer program]. http://www.lextutor.ca/vp/comp/

Coxhead, A., & Walls, R. (2012). TED Talks, vocabulary, and listening for EAP. *TESOLANZ Journal*, *20*(1), 55–67.

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes, 33*(1), 66–76.

Durbahn, M., Rodgers, M., & Peters, E. (2020). The relationship between vocabulary and viewing comprehension. *System*, *88*, 102166.

Gorsuch, G., & Taguchi, E. (2008). Repeated reading for developing reading fluency and reading comprehension: The case of EFL learners in Vietnam. *System*, *36*(2), 253–278.

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

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). Range [Computer software]. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs

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

Kaneko, M. (2014). Is the vocabulary level of the reading section of the TOEFL Internet-based test beyond the lexical level of Japanese senior high school students? *Vocabulary Learning and Instruction*. http://dx.doi.org/10.7820/vli.v03.1.kaneko

Kaneko, M. (2015). Vocabulary size required for the TOEFL iBT Listening Section. *The Language Teacher*, *39*(1), 9–14.

Kobeleva, P. P. (2012). Second language listening and unfamiliar proper names: Comprehension barrier? *RELC Journal*, *43*(1), 83–98.

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

Laufer B. (1992). How much lexis is necessary for reading comprehension? In: P.J.L Arnaud, & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 126–132). Palgrave Macmillan.

Laufer, B. (2013). Lexical thresholds for reading comprehension: What they are and how they can be used for teaching purposes. *TESOL Quarterly*, *47*(4), 867–872.

Laufer, B. (2020). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *TESOL Quarterly*, *54*(4), 1076–1085.

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.

Liu, N. & Nation, I. P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal*, *16*(1), 33-42.

Martinez, R., & Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly*, *45*(2), 267–290.

McLean, S. (2021). The coverage comprehension model, its importance to pedagogy and research, and threats to the validity with which it is operationalized. *Reading in a Foreign Language*, *33*, 126–140.

Meara, P. M., & Jones, G. (1989). *Eurocentres Vocabulary Size Test: Version E1. 1, K10, MSDOS*. Eurocentres.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines, 5*(1), 12–25.

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

Nation, I. S. P. (2012). The BNC/COCA word family lists. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, P. (2007). The four strands. *International Journal of Innovation in Language Learning and Teaching*, *1*(1), 2–13.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Heinle.

Noreillie, A. S., Kestemont, B., Heylen, K., Desmet, P., & Peters, E. (2018). Vocabulary knowledge and listening comprehension at an intermediate level in English and French as foreign languages: An approximate replication study of Stæhr (2009). *ITL-International Journal of Applied Linguistics*, *169*(1), 212–231.

Nurmukhamedov, U. (2017). Lexical coverage of TED Talks: Implications for vocabulary instruction. *TESOL Journal*, *8*(4), 768–790.

Nurmukhamedov, U., & Webb, S. (2019). Research timeline: Lexical coverage and profiling. *Language Teaching*, *52*(2), 188–200.

Pujadas, G., & Muñoz, C. (2020). Examining adolescent EFL learners' TV viewing comprehension through captions and subtitles. *Studies in Second Language Acquisition*, *42*(3), 551–575.

Qian, D. D. (1999). Assessing the roles of depth and breadth of knowledge in reading comprehension. The *Canadian Modern Language Review*, *56*(2), 282–308.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513–536.

Quaglio, P. (2009). *Television dialogue: The sitcom Friends vs. natural conversation.* John Benjamins.

Rodgers, M. P. H. & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly*, *45*(4), 689–717.

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, *50*(2), 212–226.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43.

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

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139–152.

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

Stoeckel, T., McLean, S., & Nation, P. (2020). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, 1–23.

Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based case study. *RELC Journal*, *25*(2), 34–50.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

Tegge, F. (2017). The lexical coverage of popular songs in English language teaching. *System*, 67, 87–98.

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479.

Webb, S. (2010). Pre-learning low frequency vocabulary in second language television programs. *Language Teaching Research*, *14*(4), 501–515.

Webb, S. (2016). Learning vocabulary through meaning-focused input: Replication of Elley (1989) and Liu & Nation (1985). *Language Teaching*, *49*(1), 129–140.

Webb, S.A. & Chang, A, C-S, (2012). Second language vocabulary growth. *RELC Journal*, *43*(1), 113–126.

Webb, S., & Macalister, J. (2013). Is text written for children appropriate for L2 extensive reading? *TESOL Quarterly*, *47*(2), 300–322.

Webb, S. & Nation, I. S. P. (2008). Evaluating the vocabulary load of written text. *TESOLANZ Journal*, *16*, 1–10.

Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned.* Oxford University Press.

Webb, S., & Paribakht, T. S. (2015). What is the relationship between the lexical profile of test items and performance on a standardized English proficiency test? *English for Specific Purposes, 38*, 34–43.

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

Webb, S. & Rodgers, M. P. H. (2009b). The lexical coverage of movies. *Applied Linguistics*, *30*(3), 407–427.

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics, 168*(1), 34–70.

About the Author

Stuart Webb is a Professor of Applied Linguistics at the University of Western Ontario. His articles have been published in journals such as *Applied Linguistics*, *Language Learning,* and *TESOL Quarterly*. His latest books are *The Routledge Handbook of Vocabulary Studies* and *How Vocabulary is Learned* (OUP, with Paul Nation).