

Research Report

Replacing Unknown Words from Authentic Texts

Robert F. Dilenschneider, Kathryn Zidonis Dilenschneider

Jichi Medical University
Heisei Kokusai University**Abstract**

This paper examines the use of a thesaurus to modify the unknown or difficult words that learners might encounter when they read an authentic text. First, the potential problems that learners might confront when they use a dictionary to learn several unknown words in a text are discussed. Next, word frequency levels or categories of words that make up different texts are explained. Third, studies that investigate the principle of Lexical Threshold Theory are reviewed. Last, based on previous research, pedagogical implications for using a computer thesaurus to modify the unknown words of a text are presented.

(Key words; Thesaurus, Reading Comprehension, Unknown Words)

Problems of too Many Unknown Words in Texts

If reading passages have a high density of unknown words, the time required to read the passage, the ability to learn new words, and the degree of passage comprehension for learners can be compromised. These types of obstacles can occur whenever learners need to look up several unknown words in a dictionary. For example, the first problem is that previous research has shown that it takes longer for learners to read passages when they look up unknown words in dictionaries (Lupescu & Day, 1993; Nesi & Meara, 1991). When performed numerous times, multiple look-ups for several unknown words can accumulate and cause a significant loss of time when reading a passage.

The second problem of having to look up several unknown words in reading passages concerns the difficulty of correctly recalling word spellings and meanings of unknown words. For example, previous studies have shown that if learners must learn several unknown words to understand a text, they risk losing focus and might look up the wrong word in a dictionary (Bogaards, 1998; Lupescu & Day, 1993; Tang, 1997). As a result, the benefit of using a dictionary to learn the meanings and spellings of words can be negated.

The third problem of learners having to look up several unknown words in reading passages concerns passage comprehension. Previous investigations of looking up words in a traditional dictionary while reading have indicated that this action can interfere with readers' short-term memory (Bensoussan & Laufer, 1984; Knight, 1994). Each time

learners look up an unknown word in a dictionary, their attention shifts away from comprehending passage content to focusing on word meanings. This is especially the case if learners must leave a text and are unable to read both the definition of a word and the reading passage in a side by side format.

Word Frequency Levels

To remedy the problems of too many unknown words in a text, it is important to minimize the number of unknown words learners encounter in written texts so that they can maximize their learning. One method of reducing the number of unknown words in a written text in a timely manner is to use a computer thesaurus from an Apple or Windows computer to replace low frequency words that might be unfamiliar to learners. However, in order to understand how thesauruses can be used effectively, it is first essential to recognize the different categories of words that learners encounter when reading texts.

The first category concerns high frequency words, which is comprised of the 2,000 most frequent word families (e.g., the *General Service List* [GSL] [West, 1953]) and covers approximately 68.5% of running words in spoken and written texts (Nation, 2014). If an additional 9.2% of technical words concerning the topic and subject areas belonging to the first 2,000 word families are also included, the cumulative coverage for this category of words would amount to 77.7% ($68.5 + 9.2 = 77.7\%$) (Nation, 2014). However, an analysis using the British National Corpus has indicated that the

first 2,000 words, along with proper nouns, transparent compounds, and marginal words, account for around 90% of words in an average text (Nation, 2014). Alternatively, Schmitt and Schmitt (2012) suggested that high-frequency words should be seen as the first 3,000 word families. An analysis with this expanded list, along with proper nouns, transparent compounds, and marginal words, provide around 95% coverage of an average written text.

A second category concerns mid-frequency words, which is comprised of the next 7,000 words beyond the high-frequency word list, and stretches from the third 1,000 to the ninth 1,000 word families. These words provide around an additional 9% coverage of an average text. Although these words occur less frequently than high-frequency words, they are useful to know because they largely consist of general purpose vocabulary that can help learners read texts without reference to an external resource. For instance, together with high-frequency words and proper nouns, mid-frequency words can help learners reach 98% coverage of a text (Nation, 2014).

A third category is low-frequency words, which are words beyond the 9,000 word families. Although these words make up the largest group of words, they typically account for only 1-2% of the words in a text. In spite of the fact that learning the meanings of these types of words can be beneficial, Nation (2014) suggested that, due to their specialized tendency and minimal coverage, low-frequency words be learned incidentally through reading and listening.

Not included in the previous three word categories are words such as proper nouns, transparent compounds, and marginal words. Proper nouns are nouns that denote the name of a person, place, or thing, such as *Lincoln*, *Florida*, or *Carnegie Hall*. Transparent compounds, such as *ashtray* or *aftershave*, are words, “where the meaning of the compound is transparently related to the meaning of the parts” (Nation, 2014). Marginal words, such as *aah*, *er*, *ooh*, or *ssh*, are utterances that are rarely found in written texts and are not necessarily found in dictionaries. Together, these three types of words make up about 3-4% of the words in an average text (Nation, 2014).

Lexical Threshold Theory

The principle referred to as Lexical Threshold (Laufer & Ravenhorst-Kalovski, 2010) or Lexical Threshold Theory (Prichard & Matsumoto, 2011) proposes that in order for adequate comprehension of a reading passage to occur without the use of an external resource, learners should possess a receptive knowledge of a high percentage of the words in a reading text.

One of the first studies to examine what adequate means with regard to lexical coverage of a text was conducted by Laufer (1989). In this study, 100 first-year university students enrolled in a course of English for Academic

Purposes, read a text and were given multiple-choice and open-ended questions to determine their reading comprehension score. The students were instructed to underline unknown words in the text to determine the percentage of words they knew, or their lexical coverage score. The two scores were then analyzed with the lexical coverage scores as the independent variable and reading comprehension scores as the dependent variable. Based on a passing score from the English for Academic Purposes course, adequate reading comprehension was established at 55%. A *t*-test comparing the mean scores from the reading comprehension tests showed a significant difference between students who had 95% and above lexical coverage compared to students who had 94% and below lexical coverage ($t = 8.25 > 3.46$, $p = .001$). Students familiar with 95% of the words in a text were able to achieve a passing reading comprehension score of 55% or higher while participants with lower lexical coverage failed to produce passing scores.

Laufer (1992) also examined the relationship of learners' vocabulary size with text comprehension. This study involved 92 first-year university students who completed the Vocabulary Levels Test (Nation, 1983) or Eurocentres vocabulary tests (Meara & Jones, 1990). As with the previous study, adequate comprehension was established at a similar level of 56% comprehension. The results showed that a vocabulary of the first 3,000 word families predicted a reading score of 56%. In addition, a linear regression analysis revealed that for every additional 1,000 word-frequency level, reading comprehension scores increased by 7%. For instance, a 4,000-word level predicted a reading score of 63% and a 5,000-word level predicted a reading score of 70%.

Although the reading scores were higher for learners with knowledge of lower-frequency word families compared to learners with knowledge of only higher-frequency word families, adequate comprehension was found to be significantly higher at the transition between the 2,000 to the 3,000 word frequency levels. This finding might indicate that L1 readers need to reach the 3,000 word frequency level to transfer their L1 reading strategies to L2 texts. However, the author pointed out that the relationship between reading and vocabulary size might not always be linear because as learners reach advanced vocabulary levels, improvement in reading scores likely decreases.

Hu and Nation (2000) examined the relationship between reading comprehension and text coverage by replacing low-frequency words with nonsense words. In the study, 66 adults, who were among the most proficient learners from a pre-university English course taken in an English speaking country, were divided into four groups of 16-17 people. Adequate reading comprehension was based on 12 out of 14 correct answers on a multiple-choice test and 70 out of 124 correct answers on a cued written recall test. Each group

read a 673-word story and then completed the multiple-choice and cued written recall tests.

The analyses of the four coverage groups indicated that greater text coverage led to better comprehension. For instance, for the respective multiple-choice and written recall tests, the mean scores were 6.06 and 24.60 for 80% text coverage, or one unknown word for every five words. For 90% text coverage, or one unknown word for every 10 words, the mean scores were 9.50 and 51.31. For 95% text coverage, or about one unknown word for every 20 words, the mean scores were 10.18 and 61.00. Based on this trend, the authors concluded that 98% text coverage, or about one unknown word for every 50 words in a text would provide adequate coverage for virtually all the participants in the study.

Laufer and Ravenhorst-Kalovski (2010) examined how lexical coverage, text coverage, and vocabulary size are related to reading comprehension. Their study involved a total of 745 students. In this group, 735 were college students who received scores between 75 and 133 on an English Psychometric Exam and were taking a course in English for Academic Purposes. The remaining 10 students received scores between 134 and 146 on the English Psychometric exam and thus were exempt from the English course.

The students took the English section of the Psychometric University Entrance Test to measure English reading comprehension and took a revised version of Nation's Vocabulary Levels Test (Schmitt, Schmitt, & Chapman, 2001) to measure vocabulary size. Lexical coverage was determined from the output of Tom Cobb's Vocabulary Profiler (<http://lextor.ca>), which matches the words in a text to 20 vocabulary frequency lists based on the British National Corpus (BNC). The vocabulary size, lexical coverage, and reading comprehension scores from this study are shown in Table 1.

Although the percentage of lexical coverage gradually diminished between word families, reading scores increase about 10 points. For example, between the first 2,000 word

families, the difference of the learners knowing 1,000 words and 2,000 words resulted in a difference in coverage of 9.09% ($87.67 - 78.58 = 9.09\%$), and yielded a reading score difference of 7 points ($90 - 83 = 7$). The difference of learners knowing 2,000 words and 3,000 words yielded a difference in coverage of 2.89% ($90.56 - 87.67 = 2.89\%$) and yielded a reading score difference of 12 points ($102 - 90 = 12$).

The maximum score for the reading section of the Psychometric University Entrance Test was 150. Learners who scored between 116 and 133 and had a receptive vocabulary of the first 4,000 to 5,000 words were calculated to have an overall coverage of 95.5% ($92.81 + 94.00 = 186.81 \div 2 = 93.40 + 2.1 = 95.5\%$) and required assistance to read a text. However, a small number of learners earned scores of 134 or greater. These learners were able to read without the aid of a dictionary and, therefore, were deemed to have had adequate coverage. Specifically, these learners had a receptive vocabulary size of 6,000 to 8,000 words in addition to a coverage of 2.1% for proper nouns not included in Table 1. As a result, because these learners had an overall lexical coverage between 96.5% ($94.48 + 2.1 = 96.58\%$) for a 6,000-word vocabulary and 98.40% ($96.30 + 2.1 = 98.40\%$) for an 8,000-word vocabulary, the authors concluded that adequate text coverage is around 98%.

To further investigate the role of lexical coverage and dictionary use on reading comprehension, two studies were conducted by Prichard and Matsumoto (2011). Their studies included 103 lower-intermediate to intermediate proficiency first-year Japanese university students of English. First, to explore lexical coverage, three weeks prior to reading a 650-word passage, 49 students from the control group were given a pretest in which they were asked to write definitions of 71 words from a 650-word passage in Japanese. Incorrect definitions or blank responses were subtracted from the total number of words in the reading passage (650) and then divided by 650. For example, if a participant did not know 50 of the 71 pretest words, this number was subtracted from the total number of words

Table 1. *Vocabulary Size, Lexical Coverage and Reading Comprehension* (Laufer & Ravenhorst-Kalovski, 2010)

Approximate vocabulary size	Lexical coverage	Percentile on the psychometric test	Reading score <i>M (SD)</i>	<i>N</i>
1,000	78.58	50%	83 (6.0)	109
2,000	87.67	53%	90 (7.8)	199
3,000	90.56	66%	102 (8.9)	204
4,000	92.81	72%	111 (9.4)	200
5,000	94.00	83%	122 (8.3)	23
6,000	94.48	—	—	—
7,000	95.43	91-99%	138 (4.0)	10
8,000	96.30	—	—	—

Note. Missing data not provided. Percentile on the psychometric test refers to score percentile correct for the English reading comprehension portion of the Psychometric University Entrance Test.

in the reading passage (650 word reading passage - 50 unknown words = 600 known words). This figure was then divided by 650, the total number of words in the reading passage, to find a participant's lexical coverage of that passage (600 known words ÷ 650 word reading passage = 92.3% coverage). Second, three weeks later, students read the same 650-word passage and then answered a comprehension test that consisted of eight multiple-choice questions. To indicate an understanding of the main points as well as the details of the passage, a score of 70% was used as the measure for adequate comprehension. The results revealed that comprehension scores and lexical coverage of the passage significantly correlated ($r = .29, p < .05$). Although the correlation was low, the authors stated that for the majority of learners a lexical coverage of 90% to 96% was still not enough to reach adequate comprehension because less than a quarter of the participants demonstrated 70% comprehension. Nevertheless, based on a regression line, the authors speculated that at least 97% coverage might be necessary to reach adequate comprehension.

Schmitt, Jiang, and Grabe (2011) also examined the relationship between the percentage of known vocabulary and reading comprehension. The first research question concerned the relationship between percentage of vocabulary coverage and percentage of reading coverage. The study involved 661 participants of various nationalities who ranged from high school freshmen to university graduates. On average, participants underwent over 10 years of English study and their English language proficiency ranged from intermediate to advanced. First, participants completed a 15-minute vocabulary checklist to determine how much receptive vocabulary they knew. Second, participants read a passage in which they had prior background knowledge and answered comprehension items. Third, participants read a second passage in which they lacked prior background knowledge and answered comprehension questions. The comprehension questions consisted of multiple-choice questions and 16 graphic organizer completion items in which the participants filled in partially complete information. Finally, the data were entered into a spreadsheet that calculated the total vocabulary coverage each participant had for each text.

The Spearman correlation between the variables for the percentage of vocabulary coverage of a text with reading comprehension scores of a text was .407 ($p < .001$). As a result, a linear relationship was established between the number of known words in a passage and text comprehension. For example, 90% vocabulary coverage increased comprehension to 50% and 100% vocabulary coverage increased comprehension to 75%. Based on this trend, the authors suggested that 60% reading comprehension can be obtained from 95% vocabulary coverage, and that if 70% reading comprehension is

necessary to demonstrate adequate understanding of a passage, learners should have 98-99% vocabulary coverage of a text.

Pedagogical Implications

Second language learners differ in their L2 reading experiences, L2 knowledge, and familiarity with a topic (Grabe and Stoller, 2011). How a text is organized or the time learners are afforded to read a text can also impact comprehension. Due to these factors, for learners of similar language proficiency, a text might be easy for some while difficult for others. Therefore, the threshold for learners to read texts fluently may vary and depend on factors other than vocabulary. Nonetheless, previous studies have shown the evolution of the Lexical Threshold Theory in terms of the percentage of the words that learners should know in order to adequately comprehend texts. Table 2 shows the percentage of words known in a reading passage (Text Coverage) in comparison to the comprehension percentage of a text (Comprehension Level).

Table 2. *Text Coverage and Comprehension Level Comparisons*

Study	Text Coverage	Comprehension Level
Laufer (1989)	95%	55%
Laufer (1992)	95%	56%
Hu & Nation (2000)	98%	86% * 56% **
Laufer & Ravenhorst-Kalovski (2010)	98%	89% ***
Prichard & Matsumoto (2011)	97%	70%
Schmitt, Jiang & Grabe (2011)	98%	70%

Note. *Multiple-choice tests calculated from 12 correct answers of 14 questions ($12 \div 14 = 85.7\%$) .

** Written response test calculated from 70 correct responses of 124 questions ($70 \div 124 = 56.4\%$) .

*** Entrance test calculated from 134 correct answers of 150 questions ($134 \div 150 = 89.3\%$) .

As mentioned previously, high-frequency words belonging to the first 2,000 (Nation, 2014) to 3,000 (Schmitt and Schmitt, 2012) word families, along with proper nouns, transparent compounds, and marginal words, can provide around 90% to 95% coverage of an average written text. Yet, knowledge or coverage of high-frequency words and what is deemed as adequate comprehension of a text may vary. However, because latter studies that have explored Lexical Threshold Theory have shown that the percent of comprehension for reading passages varied from 70% (Prichard & Matsumoto, 2011; Schmitt, Jiang & Grabe,

2011) to 89% (Laufer & Ravenhorst-Kalovski, 2010), it appears that if language learners know approximately 98% of the vocabulary incorporated within a typical reading passage, they will understand the majority of its content.

The use of a computer thesaurus to replace unknown words in a reading passage from a book, newspaper or magazine, might reduce the time and energy required for learners to guess the meanings of words from context or look them up in a dictionary. However, there are a couple points for language instructors to be mindful of when modifying the vocabulary of a reading passage. First, studies that have examined word frequency levels suggest that language instructors should replace the unknown mid-frequency words within a passage from the third or fourth 1,000 to the ninth 1,000 word families with high frequency words or synonyms from the first 2000 to 3,000 word families. Doing so will enable learners to read words in texts that are less common with words they are likely to already know. Second, studies that have examined Lexical Threshold Theory suggest that a high percentage of unknown words within a text should be replaced. Specifically, there should be about two unknown words for every 100 words or about five unknown words for a typical 250-word double-spaced page. This amount of coverage, however, does not necessarily ensure complete comprehension. Rather, because studies have proven that unfamiliarity with two percent of the words in a passage correlated to about 70% comprehension and being unfamiliar with five percent of the words in a passage correlated to about 60% comprehension, a higher percentage of words that might be unknown in a text may compromise learners' ability to grasp a basic or adequate comprehension of a reading passage (Schmitt, Jiang, and Grabe 2011). Therefore, due to this trend, when using a computer thesaurus to modify vocabulary from articles for classroom use, language instructors should be mindful that few words should be left unknown to learners in order to promote reading passage comprehension.

Bibliography

- 1) Bensoussan, M., & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, 7, 15-32.
- 2) Bogaards, P. (1998). Using dictionaries: Which words are looked up by foreign language learners? In B. T. S. Atkins & K. Varantola (Eds.), *Studies of dictionary use by language learners and translators* (pp. 151-157). Tübingen, Germany: Niemeyer.
- 3) Grabe, W., & Stoller, F.L. (2011). *Teaching and researching reading* (2nd ed.). New York, NY: Pearson Linguistics.
- 4) Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13, 403-430.
- 5) Knight, S. M. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *Modern Language Journal*, 78, 285-299.
- 6) 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). Clevedon, UK: Multilingual Matters.
- 7) Laufer, B. (1992). Corpus-based versus lexicographer examples in comprehension and production of new words. *EURALEX '92 Proceedings*, 71-76.
- 8) 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.
- 9) Luppescu, S., & Day, R. R. (1993). Reading dictionaries and vocabulary learning. *Language Learning*, 43, 263-287.
- 10) Meara, P. M., & Jones, G. (1990). *Eurocentres Vocabulary Size Test 10 KA*. Zurich: Eurocentres.
- 11) Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, 5, 12-25.
- 12) Nation, I. S. P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language*, 26, 1-16.
- 13) Nesi, H., & Meara, P. (1991). How using dictionaries affects performance in multiple-choice EFL tests. *Reading in a Foreign Language*, 8, 631-643.
- 14) Prichard, C., & Matsumoto, Y. (2011). The effect of lexical coverage and dictionary use on L2 reading comprehension. *Reading Matrix: An International Online Journal*, 11.
- 15) Schmitt, N., & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 1-20.
- 16) Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in text and reading comprehension. *The Modern Language Journal*. 95, 26-43.
- 17) Schmitt, N., Schmitt, D., & Chapam, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18, 55-88.
- 18) Tang, G. M. (1997). Pocket electronic dictionaries for second language learning: Help or hindrance? *TESL Canada Journal*, 15, 39-57.
- 19) West, M. (1953). *A General Service List of English Words*. London: Longman, Green & Co.

本物のテキストから不明な単語を置き換える

ロバート デイレンシュナイダー¹

キャサリン ジドニス デイレンシュナイダー²

¹自治医科大学

²平成国際大学

要 約

本研究では、シソーラスを使用して、学習者が本物のテキストを読むときに遭遇する可能性のある未知のまたは難しい言葉を修正する方法について検討します。まず、学習者が辞書を使ってテキスト中の未知語を学習する際に直面する問題について議論する。次に、単語の頻度レベルまたは異なるテキストを構成する単語のカテゴリについて説明します。第3に、語彙閾値理論の原理を研究する研究がレビューされている。最後に、以前の研究に基づいて、コンピュータシソーラスを使用してテキストの未知語を修正するための教育的含意が提示される。

(キーワード：, シソーラス, 読解, 不明な単語)