TOJELT Turkish Online Journal of English Language Teaching
A Triannual Peer-Reviewed Journal of Research in English Language Teaching
Journal Homepage: http://www.tojelt.com

Year:2017 Volume:2 Issue:3 Pages:151-155

Do You Need to Read 8,000,000 Words to Acquire 2,000 Word Families in English?: A Comment on Hill and Laufer (2003)

Jeff McQuillan
Center for Educational Development, Inc.
Senior Researcher
jeff@eslpod.com

Recommended citation: McQuillan, J. (2017). Do You Need to Read 8,000,000 Words to Acquire 2,000 Word Families in English?: A Comment on Hill and Laufer (2003). *Turkish Online Journal of English Language Teaching (TOJELT), 2*(3), 151-155.

Received:
01 June 2017
Final version:
23 July 2017
Accepted:
12 September 2017

jeff@eslpod.com

© 2017 TOJELT.
All rights reserved.

Abstract: This brief research note comments on estimates made by Hill and Laufer (2003), and repeated by other researchers, regarding the total number of words one needs to read to acquire 2,000 word families in English. Hill and Laufer (2003) only considered the acquisition of 23 previously unknown words from a text of 20,000 words. Subjects acquired five of them. From this, they concluded that readers will only acquire five new words from a text of 20,000 words. But they considered a 23-word sample, ignoring unknown words in the rest of text, under the incorrect assumption that words appearing only once in a text are never even partially acquired.

Keywords: *vocabulary, acquisition, research design, reading, ESL*

1. Introduction

Reading is an important source of vocabulary growth for both first- (L1) and second-language (L2) acquirers (Krashen, 1989; Nagy, Herman, & Anderson, 1985). This growth comes primarily through *incidental* word acquisition – that is, when readers are focused on comprehending the meaning of a text, not on building their vocabulary *per se*. Incidental vocabulary acquisition is often distinguished from *intentional* word learning, where learners are consciously attempting to learn or memorize new words.

The most common research design to measure incidental vocabulary acquisition is a "read-and-test" study, where subjects are asked to read a text that contains words unknown to them, after which they are given a surprise vocabulary test (see Krashen, 2004 for a review). Studies in L1 vocabulary acquisition have found that probability of acquiring an unknown word in English incidentally while reading is between .05 and .15 (Swanborn & de Glopper, 1999). L2 studies have produced similar estimates (e.g. Pellicer-Sánchez & Schmitt, 2010).

Some read-and-test studies employ modified texts in which all of the words are assumed known to the subjects except a small set of embedded target words. These target words are often pseudo-words (invented words) (e.g. Waring & Takaki, 2002), used to ensure that the subjects have no prior knowledge of them before reading. In such studies, *all* the target words are typically tested to measure the amount of word acquisition.

In other studies, researchers select real words that appear in a text but are thought to be unknown to the reader, based on pilot testing with a similar group of students, a pre-test given before the text is read, and/or because they are words that would only likely be encountered in

that text. Examples of this latter case include foreign words appearing in novels such as *A Clockwork Orange* (e.g. Saragi, Nation, & Meister, 1978) or *Things Fall Apart* (Pellicer-Sánchez & Schmitt, 2010). In these studies, it is more likely that a sample of the potentially unknown words is chosen to represent the larger population of unknown words.

Proper calculation of amount of word knowledge gained through incidental exposure to words while reading depends in part on distinguishing between the absolute number of words gained from the target word *sample* and the potential gains from the entire *population* of unknown words.

2. Hill and Laufer (2003): Samples, Populations, and the 8,000,000 Word Estimate

An initial error in understanding how word gains should be calculated can lead to further errors in estimating word growth from reading, as was the case with Hill and Laufer (2003). The researchers claimed that a second language reader in English would need to read 8,000,000 words in order to acquire a mere 2,000 new word families. Schmitt (2008) repeated this claim in order to bolster his argument that developing a vocabulary sufficient to read native-level texts requires a strong dose of direct instruction, since gains from incidental acquisition were too low to be counted on within a reasonable timeframe.

Here is Hill and Laufer's (2003) reasoning for their calculation:

> Studies on vocabulary acquisition from reading (without any enhancement tasks) show that pick up rates of unfamiliar words range from 1–5 words in a text of over 1,000 words (Hulstijn 1992; Knight 1994; Luppescu & Day 1993; Paribakht & Wesche 1993; Zahar et al. 2001). Similar gains occur during reading books. In Horst et al.'s (1998) experiment, an average of five words were gained from the reading of a simplified version of *The Mayor of Casterbridge*, a text of 21,000 words. Lahav (1996) conducted a study with students who read four simplified readers, each one of about 20,000 words, and found an average learning rate of 3–4 words per book. *At this rate of growth, a second language learner would have to read in excess of eight million words of texts, or about 420 novels to increase their vocabulary by 2,000 words*. This would appear to be a daunting and time consuming means of vocabulary development. It is therefore reasonable that L2 learners acquire their vocabulary not only from input, be it reading or listening, but also through word-focused activities. (p. 88, emphasis added)

How did Hill and Laufer (2003) arrive at this conclusion, and is their estimate correct?

The authors give the example of Horst, Cobb, and Meara's (1998) study of students reading a simplified reader version of *The Mayor of Casterbridge* (the Lahav study mentioned is unpublished). In Horst et al. (1998), the researchers tested students on a set of 45 words. The 45 words included eight that occurred seven or more times in the text but, because they were not part of a list of high-frequency words students studied in another aspect of their language course, were likely to be unknown to the subjects. The rest of the words were randomly selected from among low- and medium-frequency words, occurring in the text six or fewer times.

A pretest determined that the students already knew about half of the 45 target words, so the average number of new words students could have acquired on the test that was administered to them was about 23. After reading the novel, students took the post-test, which showed an average gain of around five words out of the 23 or so target words that were new to the students. The simplified novel they read contained a little more than 20,000 words. Hill and Laufer (2003) concluded therefore that you can only expect to pick up around five words for every 20,000 words you read.

Based on Horst et al.'s (1998) data, Hill and Laufer (2003) arrived at the figure 8,000,000 words that one would need to read to acquire 2,000 words: If you acquire 5 words for every 20,000 words you read, then you would need to read 400 of these 20,000 word novels, or 8,000,000 words.

The error here should be clear: Horst et al. (1998) did *not* find that only five words were acquired by students after reading a 20,000 page book, but rather that subjects got five words correct out of the *sample* of words tested. Hill and Laufer (2003) confused the population of all the unknown words in the text with the sample of words that were included in the test; Krashen, (2004) makes a similar point about this study.

Horst et al. (1998) initially estimated that the actual population of unknown word families in the text was 222. They then eliminated any word that appeared only once, which left them with 75 words, and then sampled 45 words from that list to create their test. Overall, their subjects acquired 22% of the new words they encountered, as measured on an immediate post-test.

Although the researchers thought that words that occur only once in a text were not good candidates for being acquired, other research has found that the cumulative effect of even a low rate of acquisition of single-occurring words can be substantial, in part because the majority of unknown words in a text fall into this category (McQuillan, 2016a). Pellicer-Sánchez and Schmitt (2010), for example, found that the meaning of 5% of the words that appeared only once in the text they used to measure incidental acquisition were recalled by their subjects on a post-test. Waring and Takaki's (2003) subjects recognized the meaning of 16% of the pseudo-words occurring only once on an immediate post-test similar to Horst et al.'s (1998) test.

Even if we assume that Horst et al.'s (1998) subjects already knew half of the untested words (as they knew half of the 45 tested words), and that the pick up rate was only 10%, that would leave around nine additional words acquired, effectively tripling the total number of words acquired (222 total words – 45 test words = 177 untested words, divided by 2 = 88.5 , multiplied by .10 = 8.5 words). Horst et al.'s (1998) study, then, does not provide evidence for Hill and Laufer's (2003) claim that we need to read 8,000,000 words in order to acquire 2,000 new word families.

3. Reading to Acquire 2,000 Words

The best recent estimates on the number of words one would need to read to acquire 2,000 word families in English come from a corpus analysis by Nation (2014). Table 1 summarizes Nation's estimates on the number of words one would need to acquire the 4th through the 8th most frequently occurring word families in English. Note that most adult-level texts in English can be read with vocabulary coverage of 98% at the 8,000- to 9,000-word-family level, and many popular best sellers can be read with knowledge of only 6,000- to 7,000-word families (McQuillan, 2016a).

Table 1. Number of Words One Needs to Read to Acquire the 4th through 8th Most Frequently Occurring Word Families in English

Word Family Level	Number of Words One Needs to Read	Hours @ 150 wpm	Cumulative Hours of Reading
4,000	500,000	56	56
5,000	1,000,000	112	168
6,000	1,500,000	167	335
7,000	2,000,000	222	557
8,000	2,500,000	278	835

Adapted from McQuillan (2016a), Table 1, p. 65.

Table 1 is read this way: To acquire the 1,000 word families at the 4,000-word-family level, you would need to read about 500,000 words of text (novels, magazines, newspapers, etc.). These texts would ideally be books written with 98% vocabulary coverage at the 3,000-word-family level; that is, 98% of the running words in the text would be in 1,000- to 3,000-word-family range, with only 2% at the 4,000-word-family level and above. To acquire the next 1,000 word families (the 5,000-word-family level), you would need to read 1,000,000 words of text written with all but 2% of the running text written at the 1,000- through 4,000-word-family levels, and so on. McQuillan (2016a) demonstrated that there are abundant texts written at these levels, beginning with graded readers and continuing on with juvenile, young-adult, and popular series books, including John Grisham thrillers, Agatha Christie mysteries, and the like.

Table 1 also contains time estimates of how long it would take to read this amount of text reading based on a somewhat conservative estimate of 150 words per minute (McQuillan & Krashen, 2007). For example, to go from the 4,000- to 5,000-word-family level would require around 56 hours of reading (500,000 words/150 words per minute = 3,333 minutes/60 minutes = about 56 hours).

Note that moving from the 5,000-word-family level to the 7,000-word-family level would require reading around 2,500,000 words, a much easier goal to attain than the 8,000,000 words Hill and Laufer (2003) estimated. This is something that most serious L2 students could easily manage. A student who reads 45 minutes a day in English for one year at (a relatively slow) 150 words a minute could read a total of 2,463,750 words, almost enough to reach this goal (see McQuillan, 2016b).

References

Hill, M., & Laufer, B. (2003). Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *International Review of Applied Linguistics, 41*(2), 87–106.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a foreign language*, *11*(2), 207-223.

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the Input Hypothesis. *Modern Language Journal*, *73*(4), 440-464.

Krashen, S. (2004). *The Power of Reading: Insights from the Research*. 2nd edition. Portsmouth, NH: Heinemann.

McQuillan, J. (2016a). What can readers read after grader readers? *Reading in a Foreign Language, 28*(1), 63- 78.

McQuillan, J. (2016b). Time, texts, and teaching in vocabulary acquisition: A rebuttal to Cobb (2016). *Reading in a Foreign Language*, *28*(2), 307-318.

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

Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly*, 233-253.

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.

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do *Things Fall Apart*? *Reading in a Foreign Language*, *22*(1), 31-55.

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*, *6*(2), 72-78.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research* 12(3), 329–363.

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

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*(2), 130-163.