CAN TED TALK TRANSCRIPTS SERVE AS EXTENSIVE READING MATERIAL FOR MID-FREQUENCY VOCABULARY LEARNING?

Wenhua Hsu
(whh@isu.edu.tw)

I-Shou University
No.1, Sec. 1, Syuecheng Rd., Dashu District,
Kaohsiung City 84001, Taiwan

Abstract: Schmitt and Schmitt (2014) labeled the first 4000 to 9000 word families as mid-frequency words and stressed their importance based on Nation's (2006) estimate that for adequate comprehension of a variety of authentic texts, knowledge of the first 9000 word families is necessary. Subsequent to this vocabulary goal is to determine what can be read extensively to increase vocabulary progressively since most words cannot be mastered through only one exposure. This research aimed to investigate how much TED talk transcripts input is needed to encounter most of the first 9000 word families for learning to occur. It first measured the vocabulary levels of TED talks for their potential as extensive reading material for mid-frequency word learning. The results show that TED talks reached the 5th to 6th 1000-word-family level at 98% lexical coverage. Corpus sizes of 0.3 to 4.8 million words of TED transcripts provided an average of 12+ repetitions for most of the words from the first 4th to 9th 1000 word families. The figures may serve as a reference for learners in extensive reading programs to decide how much effort they should make to read TED talk transcripts voluminously to reach a certain vocabulary goal.

Keywords: extensive reading, lexical coverage, mid-frequency words, TED talks, vocabulary levels

DOI: http://dx.doi.org/10.15639/teflinjournal.v31i2/181-203

Within English language learning, one challenge for learners is the huge number of words. In a series of tests on a variety of texts to measure how large a vocabulary is needed for adequate comprehension, Nation (2006) found that

knowledge of the most frequent 9000 word families plus proper nouns would provide 98% lexical coverage of most authentic or unsimplified texts (equivalent to two unknown words per hundred words). Using the 9000-word-family level as a vocabulary boundary, Schmitt and Schmitt (2014) labeled the first 4000 to 9000 word families as mid-frequency vocabulary as opposed to the first 3000 as high-frequency vocabulary and those after the first 9000 word families as low-frequency vocabulary. Concomitant with this long-term vocabulary goal (namely, knowledge of the first 9000 word families) is the concern regarding what English learners can read before moving to read more challenging texts beyond the first 10,000 word-family level.

Nation (2014) analyzed that controlled-vocabulary graded reader series can only provide English learners with enough input to reach up through the 4000-word-family level. In view of this, Nation and Anthony (2013) endeavored to modify some classic novels into mid-frequency readers and proposed that the vocabulary gap (4000—9000) can be bridged by reading mid-frequency readers. In agreement with Nation (2014), McQuillan (2016) put forward popular fiction series (e.g. *Harry Potter*, *Twilight*, *Hunger Games*) as an alternative of mid-frequency readers while Hsu (2019) suggested continual reading of VOA news to 6 million words to learn the most frequent 9,000 word families.

Past studies have so far looked at written text (mid-frequency readers, novels and news stories) that can provide an ample amount of input for mid-frequency vocabulary learning. However, there may be other options (for instance, spoken text) that would provide equal opportunities for meeting the first 9,000 word families enough times. To answer this question and to find an additional choice for English learners to do extensive reading, this research targeted TED talks, because TED has gathered a big fan base all over the world.

TED, a nonprofit media organization, devotes itself to spreading great ideas by hosting conferences and posting talks online for free viewing. Dating back to 1984, TED was initially conceived as a conference about Technology, Entertainment and Design. Today engaging with almost all topics, TED conferences and events are annually held around the globe and provide live streaming of talks. Out of humanitarian or innovative concerns, invited speakers use their success or fame to share their views on a certain topic from their disciplines, cultures or experience, often in a manner of storytelling. The speakers are given a maximum of 18 minutes (very few over 18 minutes) to

talk in English and the speech videos have subtitles in a cumulative of 100+ languages. At the time of doing this research, 3300+ videos are freely available on the TED website and English transcripts from these talks are mostly accessible.

Covering a great variety of topics from science to business to global issues, TED talks have been highly recommended for use in class and out of class in a range of ways, including the training of academic listening and oral presentation as well as academic spoken vocabulary development (Chang & Huang, 2015; Liu & Chen, 2019; Takaesu, 2013; Wingrove, 2017). Schmidt (2018) enumerated several examples of how TED talks have pedagogical potential for students beyond the inspiring content per se that can be gained from listening to them. For example, TED speakers may serve as a role model for English speech. Their speech styles and manners may help English learners to pick up some effective presentation skills and techniques.

Despite the fact that TED talks can be a source of language learning, yet there has been almost no attempt to explore the possibility of using TED talk transcripts as extensive reading material. As such, this study aims to examine this possible spoken text resource that English learners can use to fill the vocabulary gap between the 3000 word-family level (where low-intermediate learners leave off) and the 10,000-word-family level (where more challenging texts begin).

When implementing extensive reading, both teachers and students may want to know the amount of reading for effectiveness to occur at each stage of vocabulary development. A clear reading goal in terms of the minimum number of words needed to reach a certain vocabulary level may help students to determine how much effort they should put in to do extensive reading. This research seeks to answer the following two questions concerning TED transcripts as resources of extensive reading.

1. What is the vocabulary level of TED talks at 98% lexical coverage?
2. How much TED transcripts input do learners in extensive reading programs need in order to encounter most of the mid-frequency words often enough for learning to occur?

Vocabulary Levels

As aforementioned, TED talks offer great potential for English learning. To make full use of TED talks, it is important to know how difficult they are in

terms of vocabulary levels and which proficiency level of English learners TED talks are appropriate for to begin with.

One key to understanding vocabulary levels is the frequency of occurrence in association with the likelihood of encounter. Drawing upon the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), Nation (2017) edited 25,000 word families, which include all the affixes from levels 2 to 6 based on Bauer and Nation's (1993) criteria, and ranked them into twenty-five 1000-word-family levels according to their frequency and dispersion in the corpora. Along the BNC/COCA word-frequency scale, the first 1000 word families are the most frequent, followed by the second 1000 and the third 1000 and so on.

The reasoning behind this ranking is that higher-frequency and wider-range words stand a favorable chance of being met and then acquired than lower-frequency and narrower-range words (Nation, 2006), as repeated occurrences increase the prominence of a word and enhance the retention of knowledge of that word. For instance, the word *time* appears more frequently than *temporal* in the large corpora of British and American English. *Time* ranks at the BNC/COCA 1st 1000-word-family level as opposed to *temporal* and *epoch* at the 5th and the 8th 1000 respectively. If learners know *temporal* or *epoch*, it is highly likely that they are already familiar with the word *time*. Accordingly, compared with higher-frequency words, the more a text contains lower-frequency words, the more it may be loaded with words that are likely to be unfamiliar to learners.

Lexical Coverage Associated with Vocabulary Levels

The vocabulary level of a text can be approximated from lexical coverage first. Nation (2006) defined lexical coverage as "the percentage of running words in the text known by the reader" (p. 61). Language learners depend on vocabulary knowledge as their first resource to decode the meaning of a text. As the density of unknown words increases, the degree of comprehension decreases. Although knowing 100% of the words of a text (i.e., 100% lexical coverage) does not warrant 100% comprehension, higher lexical coverage is better than lower coverage (Schmitt et al., 2011) in terms of a higher probability of having a good degree of comprehension. For this reason, lexical coverage (percentage of known words) has often been considered as a good

indicator of whether a text is likely to be adequately understood (Webb & Nation, 2013).

Because reading involves degrees of comprehension, researchers have diverged in lexical coverage percentage. In a series of experiments on coverage percentage, Laufer (1989, 1992) found that for learners to be able to gain reasonable comprehension of a text, it is necessary for them to know at least 95% of the total words of that text. In later research, Laufer and Ravenhorst-Kalovski (2010) suggested two coverage degrees, 95% and 98% as the possible lower and upper thresholds over which learners are likely to gain successful comprehension. Hirsh and Nation (1992) advocated that to read for pleasure, learners need to get a grip of 98% of the words in a text. Similarly, Hu and Nation (2000) upheld that 98% lexical coverage is necessary for unassisted reading. They further pointed out that having vocabulary reaching only 80% coverage of a text (i.e. meeting one unknown word per five words), one would not be able to read the text effectively. In a sequence of tests on lexical coverage, Nation (2006) concluded that 98% coverage is the lexical threshold for adequate comprehension of a written text and for ideal guessing of words from context.

In a nutshell, past studies on lexical coverage have consistently used 95% or 98% as a benchmark. Determining how large vocabulary is needed for comprehension of a text can be approached from the lexical coverage predetermined (95% or 98%) and then by counting from the first 1000 the number of the ranked BNC/COCA 1000-word-family lists needed until an accumulation of lexical coverage reaches 95% or 98%. At the same time, the vocabulary level of a text at 95% or 98% coverage can be measured based on which 1000-word-family level is the last one being added when the cumulative coverage, counted from the coverage of the first 1000, arrives at 95% or 98%.

Past Studies on Lexical Profiling of TED Talks

Despite a rather small-scaled corpus, Nurmukhamedov and Sadler's (2011) research was one of the earliest studies to examine the lexical coverage of TED talks. They examined the lexical profiling of a 221-word section of the speech entitled 'Schools kill creativity' using the General Service List of English words (GSL) (West, 1953) and the Academic Word List (AWL) (Coxhead, 2000) as base word lists for analysis. Results showed that the GSL covered 90% of the words in that section of speech and the AWL provided 5%

coverage with the remaining 5% of the words neither belonging to the GSL nor the AWL.

Wang (2012) built a larger TED talk corpus than Nurmukhamedov and Sadler's (2011). She gathered 10 TED talks under 20 minutes in relation to three topic areas (business, global issues and science & technology) totaling 80,885 words and compared the vocabulary of TED talks with that used in a corpus of 40 long lectures and 10 seminars on physical sciences and social sciences from the British Academic Spoken Corpus (BASE) (developed at the Universities of Warwick and Reading), totaling 643,649 words. Rather than using the GSL and the AWL, Wang (2012) drew on the BNC 1000-word-family lists (Nation, 2006) to measure the lexical coverage of the two corpora. The data demonstrated that the first BNC 2000 and 3000 word families made up 91% and 93% of the total words of the TED talks respectively as opposed to their 89% and 90.65% coverage in the BASE in turn. Compared with academic spoken English, TED talks may be easier to begin with for English learners with a mastery of 2000-3000 word families.

In a different line of research, Wolfe (2015) collected 1,790 TED talk transcripts with a total of 3,868,390 tokens and therein created a 421-word-family TED talks Word List (TWL) in a bid to facilitate English learning from TED talks. The TWL provided another 2.7% coverage point after the GSL and the AWL. The GSL accounted for 88.03% of the words in TED talks while the AWL covered 3.73%. Though Wolfe's (2015) corpus size was much larger than Nurmukhamedov and Sadler's (2011), there was not much disparity between the two in terms of the coverage of the GSL and the AWL in TED talks.

As shown in Wolfe's (2015) as well as Nurmukhamedov and Sadler's (2011) data results, knowledge of the GSL would enable one to understand 88.03% or 90% of the words of TED talks. Earlier studies reported that the GSL provided coverage of 78% to 92% of all sorts of written text, averaging 82% coverage (Hirsh & Nation, 1992; Sutarsyah et al., 1994). When tested on Coxhead's (2000) four academic domains of texts (arts, commerce, law and science), the GSL offered the lexical coverage of 77.4%, 76.8%, 79.1% and 70.7% respectively. It is worth noting here that the nearly 2000-word-family GSL (West, 1953) has long been considered as basic vocabulary of English, since it contains the most frequent general service words. As far as the difficulty of TED talks – as measured by the coverage of the GSL– is concerned, TED talks may be moderately challenging for English learners with a goal to learn mid-frequency words if compared with the above-mentioned

data figures (the GSL coverage in TED talks 88.03% to 90% versus 70.7% to 92% in various written texts). This gives a beacon of hope for low-intermediate English learners using TED transcripts as extensive reading material for mid-frequency vocabulary learning.

Targeting short talks (4 to 6 minutes in length), Coxhead and Walls (2012) compiled a 43,656-word corpus of 60 TED talk transcripts across six topic areas (business, design, entertainment, global issues, science, and technology) with each containing 10 talks to examine the vocabulary load of each domain and the coverage of the AWL in the TED talks. Through the RANGE program (Heatley, Nation & Coxhead, 2004) installed with the BNC word-frequency lists, their data showed that the first 4000 word families and the first 8000 to 9000 word families plus proper nouns were necessary to reach 95% and 98% coverage respectively. The coverage over five of the six topic areas by the first BNC 3000 word families was slightly over 92%, except that the coverage of the technology TED talks was 90.02%. Compared with around 10% coverage in written academic texts (Chen & Ge, 2007; Coxhead, 2000; Li & Qian, 2010), the AWL coverage in TED talks was only 3.90%, which is in line with Wolfe's (2015) as well as Nurmukhamedov and Sadler's (2011) findings (3.73% and 5% respectively). In terms of the AWL coverage, both TED talks and newspapers are very similar, with the former being 3.9% and the latter 4% (Chung & Nation, 2003; Coxhead, 2000). Moreover, Coxhead and Wall (2012) reported that at 98% coverage, the vocabulary level of TED talks has reached the first 8000 to 9000 word families, closer to that of written texts than that of spoken texts (e.g. the first 8000 to 9000 word families for novels, newspapers and academic texts (Nation, 2006) versus the first 6000 to 7000 word families for movies (Webb & Rodgers, 2009).

In a similar vein, Nurmukhamedov (2017) also sought to measure the vocabulary size necessary for adequate comprehension of TED talks and examined whether different registers of TED talks change the vocabulary demand based on a lexical profiling approach. He compiled a corpus of 400 TED transcripts, which were claimed to be more representative and better balanced than Coxhead and Walls's (2012). Nurmukhamedov (2017) concluded that to get a grip on 95% and 98% of the words of TED talks, the first 4000 word families and 8000 word families in turn plus proper nouns and marginal words would be necessary. Despite the larger TED talk corpus, the result was not much different from the finding by Coxhead and Walls (2012) concerning the vocabulary demand.

Since TED talks have reached the vocabulary levels of the first 8000 to 9000 word families at 98% coverage, mid-frequency words (4000 to 9000) do feature in TED talks. TED transcripts may provide a solid resource from which one can build lexical knowledge in the mid-frequency bands.

The Amount of Reading Input Necessary for Acquiring Mid-Frequency Vocabulary

To solve the debate of whether it is possible to learn enough vocabulary solely through reading, Nation (2014) used corpora of various compositions to measure how much reading input is needed to gain enough repetition of the first 9000 word families for learning to occur. Nation (2014) presumed that to have an opportunity of acquiring an unknown word, one must encounter it in text enough times. Referring to previous research on incidental learning and repetition (Vidal, 2011; Waring & Takaki, 2003; Webb, 2007), Nation (2014) conjectured that 12 repetitions would be enough to allow for a chance to learn a new word and used 12 repetitions as the cut-off point for subsequent series of calculation.

Table 1. Nation's (2014) Recommended Amount of Novels Input Needed to Encounter Most of the Words from the 2nd to 9th 1000-Word-Family Levels

Learners' vocabulary level	Necessary amount to read (in words)	To encounter most of the words at this 1000-word-family level	Cumulative amount
1st 1000	200,000	2nd 1000	200,000
2nd 1000	300,000	3rd 1000	500,000
3rd 1000	500,000	4th 1000	1,000,000
4th 1000	1,000,000	5th 1000	2,000,000
5th 1000	1,500,000	6th 1000	3,500,000
6th 1000	2,000,000	7th 1000	5,500,000
7th 1000	2,500,000	8th 1000	8,000,000
8th 1000	3,000,000	9th 1000	11,000,000

As illustrated in Table 1, Nation (2014) compared word learning to climbing up a sequence of staged steps. He put forward this notion that a text which can be read at 98% coverage at a certain 1000-word-family level can be used to help learners to acquire words in the next 1000 level. In his chosen

corpus of 25 novels from Project Gutenberg (http://www.gutenberg.org/), Nation (2014) estimated that English learners with knowledge of the first 1000 word families would need to read about 200,000 words of novels in order to encounter 800+ word families at the 2nd 1000-word-family level enough times to have a chance of acquiring them. Learners with knowledge of the first 2000 word families would need to read approximately 300,000 words to meet 800+ word families at the 3rd 1000 level often enough. By the same token, to learn the 4th 1000 word families, learners would need to read 500,000 words in order to gain 12+ encounters with most of the words at that level. From the 5th 1000 onwards, the increase in the amount of reading is half a million words for the learning of the next 1000 word families (see Table 1). Finally, to meet most of the words at the 9th 1000 level sufficient times, learners with knowledge of the first 8000 word families would need to read 3 million words. If learners with the knowledge of the first 1000 word families follow this reading scheme, they would have read an accumulation of 11 million words of novels by the time their vocabulary size increases to the first 9000 word families.

METHOD

The TED Talk Corpus

The present study did not use a web scraper to automatically retrieve TED webpage content in Rich Text Format (RTF). To avoid unreadable messy code via web scraping, which may take more time to tidy up the transcripts later, the researcher paid twenty-four English-majored students to help instead. During the data collection, they copied a total of 3,776 English transcripts from the TED website and then pasted each to a Word document and saved it as a plain text file in UTF-8.

The TED website divides speech themes into six major topics: technology, entertainment, design, business, science, and global issues. However, each TED talk may involve several inter-related subtopics and hence has multiple topic tags. The TED website lists all topics from A (e.g. activism) to T-Z (e.g. youth and 3D printing). This means that some talks may appear in more than one of these topical categories. When calculating the total tokens of the corpus, the researcher first identified repeated talks with the aid of the alphabetically sorting function in Excel and made sure the repeated TED talks were not repeatedly counted.

Since the topical categorization of a TED talk may not be clear-cut, this research followed the website's six-topic classification and put the talks into six sub-corpora. Table 2 summarizes the number of English transcripts and the word count of each sub-corpus.

Table 2. The TED Talk Corpus

Topic	Number of talks	Word count	Average length per talk
Technology	935	1,900,096	2,032
Entertainment	375	651,393	1,737
Design	532	1,025,276	1,927
Business	447	980,748	2,194
Science	908	1,728,406	1,904
Global Issues	579	1,288,040	2,225
Total talks with overlap	3,776	7,573,959	2,006
Total talks without overlap	2,630	5,286,303	2,010

Note: There was an overlap of 1,146 talks across six main topics.

The Instruments and Data Processing

Each talk transcript contains non-spoken text, identified as words in parentheses or in brackets. They include reactions from the audience marked by parentheses as well as transcribed text from the speaker's PowerPoint slides marked by brackets. The researcher decided to remove these two non-spoken words, *applause* and l*aughter*, since they appear at least once per talk, either at the end of each talk or throughout the talk, leading to such a high frequency that having these words included would overestimate lexical coverage and underestimate vocabulary demands. In addition, two expressions, *thank you* and *thank you very much,* were decided to be deleted. Though they do not appear in parentheses or in brackets, they occur at the end of almost every talk or after applause. Their very high frequencies may inflate lexical coverage as well. To avoid overestimate, the researcher used the text editing software TextMate 2.0 (https://macromates.com) to find and replace these words or expressions across files with a blank space. They were summarily deleted.

However, the researcher kept the unspoken PPT text intact. Research on TED talks with a focus on listening may ignore the transcribed words from PPT slides since they are not spoken. However, when TED transcripts are used for reading purposes, transcribed PPT words should be included as they come along with the transcripts which they appear in.

Following Nurmukhamedov (2017), Coxhead and Wall (2012) as well as Wang (2012), this research also adopted the RANGE program (Heatley, Nation, & Coxhead, 2004) by installing the 25 ranked BNC/COCA 1000-word-family lists to calculate lexical coverage. Apart from the BNC/COCA 25,000 word families, the present study used three additional lists in the RANGE (i.e. Basewrd31 for proper nouns, Basewrd32 for marginal words and Basewrd33 for transparent compounds) to provide the lexical profiling of TED talks. The three word lists were considered as having a minimal reading burden. English learners whose vocabulary goal is to learn the first 4000 to 9000 word families should recognize the name of a place or a person from its spelling without much effort. Marginal words such as spoken interjections and exclamations (huh, erm and ooh) would not cause many difficulties in reading comprehension. It is also not difficult to infer the meaning of transparent compounds from their constituent words if they are known to learners. A separate listing of transparent compounds would prevent repetitive counting since their component words have already been included in the BNC/COCA 25,000 word families.

In short, doing without proper nouns, marginal words and transparent compounds would have overjudged the vocabulary levels of TED talks. Therefore, the coverage percentage of these three word lists were added until the cumulative coverage arrived at 95% or 98%.

FINDINGS AND DISCUSSION

The Vocabulary Levels of TED Talks across Six Main Topic Areas

Table 3 provides a snapshot of the lexical coverage of each of the BNC/COCA 1000-word-family lists in the TED talks and the vocabulary levels at 95% and 98% coverage.

Table 3. Lexical Coverage at Each of the BNC/COCA Base Word Lists Plus Three Additional Word Lists in the TED Talks across Six Main Topics

Word list	Technology	Entertainment	Design	Business	Science	Global Issues	Whole
Proper nouns	1.12%	1.73%	1.19%	1.3%	0.96%	1.76%	1.27%
Marginal words	0.09%	0.18%	0.10%	0.09%	0.07%	0.09%	0.10%
Compounds	0.35%	0.36%	0.38%	0.34%	0.32%	0.32%	0.34%
1st—2nd 1000	89.90%	90.52%	90.37%	90.48%	89.11%	90.11%	89.94%
3rd 1000	4.32%	3.10%	3.86%	4.25%	4.60%	4.21%	4.19%
Cumulative % at 3rd 1000	95.78%*	95.89%*	95.90%*	96.46%*	95.06%*	96.49%*	95.84%*
4th 1000	1.34%	1.24%	1.31%	1.10%	1.55%	1.09%	1.30%
5th 1000	0.86%	0.89%	0.87%	0.71%	1.01%	0.80%	0.87%
Cumulative % at 5th 1000	97.98%	98.02%**	98.08%**	98.27%**	97.62%	98.38%**	98.01%**
6th 1000	0.40%	0.42%	0.41%	0.35%	0.56%	0.37%	0.44%
Cumulative % at 6th 1000	98.38%**	98.44%	98.49%	98.62%	98.18%**	98.75%	98.45%
7th 1000	0.34%	0.32%	0.39%	0.31%	0.49%	0.28%	0.36%
8th 1000	0.30%	0.28%	0.31%	0.27%	0.39%	0.24%	0.27%
9th 1000	0.25%	0.19%	0.15%	0.22%	0.25%	0.15%	0.19%
10th 1000	0.16%	0.12%	0.11%	0.14%	0.14%	0.09%	0.14%
11th –25th 1000	0.57%	0.65%	0.55%	0.44%	0.55%	0.49%	0.59%

Note: The symbol * denotes reaching 95% coverage including proper nouns, marginal words and transparent compounds, while ** means reaching 98% plus proper nouns and so on. The right column 'Whole' means the six corpora as a whole.

Previous studies on TED talks reported that knowledge of the GSL would provide 88.03% or 90% text coverage (Nurmukhamedov & Sadler, 2011;

Wolfe, 2015), while knowing the BNC 2000 would help to master 91% of the words of TED talks (Wang, 2012). In Table 3, the coverage of the first BNC/COCA 2000 word families across six topics ranges from 89.11% to 90.52%. This shows the relative significance of knowing the most frequent 2000 word families, either the GSL or the BNC 2000 or the first BNC/COCA 2000 as a prerequisite to approaching TED talks. Highly-motivated upper-beginners with knowledge of the first 2000 word families (providing approximately 90% lexical coverage of TED talks) may be encouraged to give reading TED manuscripts a try, although they may encounter one unknown word in every line of text.

As shown in Table 3, there is not much variation in vocabulary levels of TED talks across six main topic areas. At 95% coverage, either in each corpus or in the six corpora as a whole, TED talks consistently converge at the 3000-word-family level. At 98% coverage, TED talks of different topics are mostly at the 5000-word-family level, except science and technology both stretching to the 6000-word-family level. In other words, reading TED manuscripts in relation to science or technology is the most vocabulary-demanding. Low-intermediate learners with knowledge of the first 3000 word families (providing 95% coverage) may be encouraged to read TED manuscripts, because they may not be a formidable task as we have expected. Namely, at 95% coverage, five unknown words in every 100 words of text may be tolerated before they interfere with comprehension. When learners' vocabulary has developed to the first 6000 word families, they may feel at equal ease when reading TED manuscripts of any topic, in terms of the frequency of guessing unknown words or consulting a dictionary.

As opposed to the present result that the first 3000 and 5000—6000 word families provide 95% and 98% coverage respectively, Coxhead and Walls (2012) as well as Nurmukhamedov (2017) registered a higher vocabulary demand for adequate comprehension of TED talks, with the first 4000 word families needed for 95% coverage and the first 8000—9000 word families for 98% coverage. The reason for this discrepancy may be due to the fewer additional lists and the BNC word-family lists utilized in Coxhead and Walls (2012) as well as in Nurmukhamedov (2017). This research adopted the more updated BNC/COCA word-family lists and three additional lists (proper nouns, marginal words and transparent compounds) to analyze the vocabulary level of TED talks. In contrast, Nurmukhamedov (2017) included only two lists of proper nouns and marginal words to calculate lexical coverage, while Coxhead

and Walls (2012) merely included the proper noun list, thus resulting in a higher vocabulary level and a larger vocabulary demand. The present data shows that a smaller vocabulary may suffice for good comprehension of TED talks. However, the result also implies that in order to encounter most of the first 9000 word families, learners may need to read a lot more TED manuscripts, since they only reach the first 5000—6000 word-family level at 98% coverage rather than the first 8000—9000 word-family level as reported in the literature.

Amount of TED Talks Input Needed to Encounter Most of the Mid-Frequency Words

In line with Nation's (2014) cut-off point, the present research set the threshold at 12+ repetitions of 800+ word families from each of the 4th to 9th 1000-word-family levels. Table 4 incorporates the data from Table 2 concerning the corpus size retrieved from TED talks across six main topics and the information from Table 3 about the vocabulary level of each of the six corpora at 98% coverage into the new results showing the number of word families with 12+ occurrences at each 1000-word-family level.

Since all the six topical TED corpora had over half a million words, the researcher retrieved 0.5 million words at random from each corpus as a pilot study. The pilot results showed that irrespective of any topic domain, learners may encounter 900+ word families at the 4th 1000 level twelve times or more if they keep reading TED manuscripts up to half a million words.

Table 4. Number of Word Families Occurring 12+ Times at Each 1000-Word-Family Level

	Technology talks	Entertainment talks	Design talks	Business talks	Science talks	Global Issues talks
Word count	1,900,096	651,393	1,025,276	980,748	1,728,406	1,288,040
Vocab level at 98% coverage	6000	5000	5000	5000	6000	5000
4th 1000	*983*	*918*	*941*	*940*	*976*	*974*
5th 1000	*931*	*804*	*856*	*844*	*923*	*914*

	Technology talks	Entertainment talks	Design talks	Business talks	Science talks	Global Issues talks
6th 1000	*867*	696	764	735	*865*	799
7th 1000	769	556	652	605	765	693
8th 1000	685	454	555	522	692	579
9th 1000	584	412	473	383	595	479

Note: The symbol * * denotes that at this corpus size, the goal of 12+ encounters with 800+ word families at a given 1000-word-family level has been achieved. 'Vocab' means vocabulary.

At the time of doing this research, there were only 651,393 words of entertainment talks available. However, Table 4 shows that at this amount, learners can encounter 804 word families at the 5th 1000-word-family level enough times. Nation (2014) (see Table 1) asserted that when learners reach the 4th 1000-word-family level, they would need to read at least one million words of text at the 4th 1000 level at 98% coverage to learn most of the 5th 1000 word families. Counter to Nation's (2014) estimate, the above result reveals that learners may not need this much to reach the goal. To test the other five corpora, a random retrieval of 0.6 and 0.7 million words (the nearest whole number, rounded down and rounded up from 651,393) from each corpus was performed repeatedly. Results validated that learners would meet 800+ word families at the 5th 1000 level 12+ times if they continue to read TED manuscripts of any topic to the amount of 0.6 million words.

In view of fewer than one million words needed for meeting most of the 5th 1000 word families, the researcher suspected that learners may not need as many as 0.5 million words of input to gain 12+ encounters with 800+ word families at the 4th 1000 level based on Nation (2014) recommended amount of reading. After repeated trials of each corpus, 0.3 million words of TED transcripts would suffice to attain this goal.

As for the learning of the 6th 1000 word families, reading as many as 1.5 million words would be needed, as per Nation (2014). However, Table 4 displays that the 1,288,040-word global issue talks would provide almost enough input to meet nearly 800 word families at the 6th 1000 level (799 word families are very close to the threshold of 800 word families). Therefore, a series of retrieval of 1.3 million words, 1.4 million words and 1.5 million words was performed on the science and technology talk corpora, since only

these two corpora contained over 1.5 million words. The result verified that to encounter 800+ word families at the 6th 1000 level 12+ times, learners with a mastery of the first 5000 word families would need to read 1.3 million words of TED science or technology transcripts.

It is worth noting here that TED talks only reach the 5th to 6th 1000-word-family levels at 98% coverage (see back Table 3). As discussed above, continual reading of TED talks of the same topic area from 0.3 million to 0.6 million to 1.3 million words would theoretically provide enough input to encounter the majority of the word families from the 4th to the 5th to the 6th 1000 levels respectively.

Nevertheless, if learners expect to increase their vocabulary to 9000 word families, they would need to read a mixture of different topical TED talks to have enough input and a better chance of acquiring a wide variety of words up through the 9th 1000 word families (Nation, 2014). Based on this principle, the researcher used Nation's (2014) recommendation on the amount needed for meeting most of the 7th 1000 word families, i.e. 2 million words as a starting point for a set of trials.

Approximately 333,334 words were retrieved from each corpus, making a total of 2 million words. At this amount, the goal of 800+ word families from the 7th 1000 occurring 12+ times was not met. Therefore, another 0.1 million words retrieved equally from each corpus were added to the 2 million words. The incremental amount was 0.1 million words each time until the threshold was reached. The same procedures were also applied to the 8th and 9th 1000 word families.

Table 5 summarizes the amount of TED talks input needed to gain 12+ encounters with 800+ word families from each of the 4th to 9th 1000 word families.

As shown in Table 5, low-intermediate learners with knowledge of the first 3000 word families would need to read about 0.3 million words of TED talks in order to encounter 800+ word families at the 4th 1000 level twelve times or more to have an opportunity of acquiring them. They may get all 0.3 million words of TED transcripts from any of the six topical corpora since each corpus had over 0.3 million words available.

Likewise, learners with knowledge of the first 4000 word families would need to read approximately 0.6 million words to meet 800+ word families at the 5th 1000 level often enough. To learn the 6th 1000, learners would need to read 1.3 million words to gain an average of 12+ encounters with most of the

words at that level. They may either stick to the reading of science or technology manuscripts from the very beginning since only these two corpora contained this much amount or they can read transcripts of various topics up to 1.3 million words.

Table 5. Amount of TED Talks Input Needed to Gain 12+ Encounters with 800+ Word Families from Each of the 4th to 9th 1000 Word Families

BNC/COCA 1,000-word-family levels	Nation's (2014) recommended amount of novels input (in words)	Approximate amount of TED talks input needed (in words)	Corpus retrieved	Number of word families occurring 12+ times at this 1000-word-family level
4th 1000	0.5 million (500,000)	0.3 million (300,000)	Technology	800
			Entertainment	805
			Design	804
			Business	812
			Science	801
			Global issues	807
5th 1000	1 million (1,000,000)	0.6 million (600,000)	Technology	806
			Entertainment	802
			Design	805
			Business	800
			Science	808
			Global issues	803
6th 1000	1.5 million	1.3 million	Technology	800
			Science	806
7th 1000	2 million	2.4 million	6 corpora equally retrieved	800
8th 1000	2.5 million	3.6 million	6 corpora equally retrieved	800
9th 1000	3 million	4.8 million	6 corpora equally retrieved	800

From the 7th 1000 onwards, the increase in the amount of reading is 1.2 million words for the learning of the next 1000 word families. Specifically, to meet most of the words at the 7th, 8th and 9th 1000 level sufficient times, learners with knowledge of the first 6000, 7000 and 8000 word families would need to read 2.4 million, 3.6 million and 4.8 million words respectively, which were a lot more than Nation's (2014) recommended amount, namely reading 2 million, 2.5 million and 3 million words of novels for the learning of the 7th, 8th and 9th 1000 word families in turn.

The results may be attributed to the fact that the vocabulary level of TED transcripts only reaches the 5th to 6th 1000 level at 98% coverage while that of Nation's (2014) data set extends to the 9th 1000 level or beyond. However, after leaners have developed their vocabulary capacity to the first 5000 to 6000 word families, reading TED transcripts at 98% lexical coverage is likely to become easier than reading classic novels or more challenging texts that go beyond the 9th 1000-word-family level. If learners keep reading TED transcripts, eventually they would meet most of the first 9000 word families enough times to have a reasonable opportunity of acquiring them.

To summarize, in addressing Research Question 2, 'How much TED transcript input do learners need in order to encounter most of the mid-frequency words often enough for learning to occur?', the present results demonstrate that to meet most of the first 9,000 word families enough times, learners would need to read 4.8 million words of TED talk transcripts at the minimum.

Pedagogical Implications

Although knowledge of the first 3000 word families would provide 95% lexical coverage (i.e. 5 unknown words per hundred words), English learners at this proficiency level may still find it a challenge to read TED talk transcripts in terms of 100 unknown words per 2000-word TED talk (see Table 2 for the average length per TED talk being around 2000 words).

At first glance, reading a 2000-word transcript with 100 unknown words may be a daunting task. However, it is not necessary to look up every unknown word while reading. Learners can often proceed to read further on, because the context of an unknown word and its later recurrence in the text may give them some clues about its meaning. Over time with the increase of vocabulary to a level of attaining 98% lexical coverage, successful guessing from context will

become easier to achieve due to fewer unknown words compared with 95% coverage.

Moreover, there are some other approaches which can reduce the frequency of dictionary lookup. When learners self-select TED transcripts to read, they are advised to choose talks covering similar subject domains at the initial stage of extensive reading. By reading transcripts of similar topics, learners can cut down the word learning load as they proceed, since similar talks contain topic-related keywords which may occur across these talks. Also, similar topical talks would offer more repetitions of keywords and therefore provide learners with a favorable condition for word learning.

Apart from English transcripts, many TED talks have translations of 100+ languages available. Choosing mother-tongue translations to pre-read may be a good support to reading their corresponding English transcripts later. With background knowledge from reading mother-tongue translations beforehand, learners may find it less difficult to guess unknown words from context when reading their English versions.

Furthermore, though not being directly relevant to this research focus, learners can read transcripts while viewing and listening to TED talks to improve their comprehension. Chang (2009) provided some evidence that learners process spoken text more effectively with the mode of reading while listening compared with listening only. Similarly, Renandya and Jacobs (2016) have also suggested implementing extensive reading and extensive listening together so that learners can benefit from input-based learning.

When learners make a resolution to do extensive reading, they may have little idea of the extent to which they should keep reading before effectiveness occurs. A concrete number in terms of the necessary amount of reading would help learners to decide how much effort they should make to read TED transcripts voluminously. Table 5 lists a set of possible amounts of TED transcripts input learners may need to have a reasonable chance of acquiring most of the 4^{th} to 9^{th} 1000 word families, contingent on their English proficiency level at which they begin their extensive reading.

To put the recommended amount of reading into perspective, Nation (2014) delineated the time needed to carry out a reading scheme for approaching a certain vocabulary level based on several assumptions. A case in point is that he estimated that at the speed of 150 words per minute, it would take learners 333 hours to read 3 million words suppose the reading material is at the right level where no more than 2% of the words are unknown. At one

hour per day, this means approximately one year of reading, which is very doable for motivated learners.

If Nation's (2014) assumptions behind this computation are correct, we can adopt a more conservative estimate by setting 100 words per minute in lieu of 150 words as a calculation basis for the present research. At the speed of 100 words per minute for extensive reading, leaners would need to spend 800 hours reading 4.8 million words of TED talk transcripts. At one hour per day, this represents less than 3 years of reading, very feasible within four years of college entry. Or if learners can read three TED talk transcripts a day, five days a week and 40 weeks a year at least, they can also complete this feat during their college years. Admittedly, reading goals in terms of completion time will differ subject to learners' proficiency level, reading speed and the time allowed every day. This issue is worth attending to but beyond the focus of this research.

Lastly, this research is limited by the fact that the TED talk databank is continually growing. However, undoubtedly, English learners can get sufficient input from TED talk transcripts to learn most of the mid-frequency words (the first 4000 to 9000 word families) through extensive reading.

CONCLUSIONS

English learners can get sufficient input from TED transcripts to learn most of the mid-frequency words through extensive reading. Although the vocabulary use of TED talks only reaches the 5th to 6th 1,000-word-family level at 98% lexical coverage, continual reading of TED transcripts up to 4.8 million words will still offer English learners a good opportunity of meeting most of the first 9000 word families enough times for learning to occur. The figures may serve as a reference for extensive reading practitioners as well as a goal for learners in extensive reading programs who are concerned with mid-frequency vocabulary learning.

REFERENCES

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253–279.

Chang, C.-S. (2009). Gains to L2 listeners from reading while listening vs. listening only in comprehending short stories. *System, 37*(4), 652–663.

Chang, Y.-J., & Huang, H.-T. (2015). Exploring TED talks as a pedagogical resource for oral presentations: A corpus-based move analysis. *English Teaching & Learning, 39*(4), 29-62.

Chen, Q. & Ge, G. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes, 26*(4)*,* 502–514.

Chung, T. M., & Nation, I. S. P. (2003). Technical vocabulary in specialized texts. *Reading in a Foreign Language, 15(2)*, 103–116.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Coxhead, A., & Walls, R. (2012). TED Talks, vocabulary, and listening for EAP. *TESOL ANZ Journal*, *20*(1), 55–65.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2004). *Range (Version 1.32.) [Computer software]*. Retrieved from https://www.victoria.ac.nz/lals/about/staff/paul-nation

Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8*(2), 689–96.

Hsu, W. (2019). Voice of America (VOA) news as voluminous reading material for mid-frequency vocabulary learning. *RELC Journal, 50*(3), 408-421.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren, & M. Nordman (Eds.), *Special language: From human 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. Bejoing (Eds.), *Vocabulary and applied linguistics* (pp. 129-132). Macmillan.

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*(1), 15–30.

Li, Y. & Qian, D. D. (2010). Profiling the Academic Word List (AWL) in a financial corpus. *System, 38*(3), 402–411.

Liu, C.-Y., & Chen, H. H.-J. (2019). Academic spoken vocabulary in TED talks: Implications for academic listening. *English Teaching & Learning, 43*(4), 353-368.

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

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59–82.

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

Nation, I. S. P. (2017). *The BNC/COCA Level 6 word family lists (Version 1.0.0) [Data file]*. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Nation, I. S. P., & Anthony, L. (2013). Mid-frequency readers. *Journal of Extensive Reading, 1*, 5–16.

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

Nurmukhamedov, U., & Sadler, R. (2011). Podcasts in four categories: Applications to language learning. In B. Facer & M. Abdous (Eds.), *Academic podcasting and mobile assisted language learning* (pp. 176–195). Book News.

Renandya, W. A., & Jacobs, G. M. (2016). Extensive reading and listening in the language classrooms. In W. A. Renandya & H. P. Widodo (Eds.), *English language teaching today: Linking theory and practice* (pp. 97-110). Springer International Publishing AG.

Schmidt, A. (2018). *How much does your TED talk? Vocabulary coverage and TED talks*. Retrieved from https://www.eltresearchbites.com/201805-how-much-does-your-ted-talk-vocabulary-coverage-and-ted-talks/

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

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484-503.

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.

Takaesu, A. (2013). TED Talks as an extensive listening resource for EAP students. *Language Education in Asia, 4*, 150–162.

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning, 61*(1), 219–258.

Wang, Y. (2012). An exploration of vocabulary knowledge in English short talks: A corpus driven approach. *International Journal of English Linguistics, 2*(4), 33–43.

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.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65.

Webb, S., & Nation, P. (2013). Computer-assisted vocabulary load analysis. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–10). Wiley-Blackwell.

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

West, M. (1953). *A general service list of English words*. Longman, Green and Co.

Wingrove, P. (2017). How suitable are TED talks for academic listening? *Journal of English for Academic Purposes, 30*, 79-95.

Wolfe, J. D. (2015). *The TED word list: An analysis of TED Talks to benefit ESL teachers and learners*. (Unpublished Master's thesis, Royal Roads University, Colwood, British Columbia, Canada). Retrieved from https://viurrspace.ca/handle/10170/864.