Article

RELC Journal
1–15

DOI: 10.1177/0033688220979315
journals.sagepub.com/home/rel
SAGE

Corpus-Based Vocabulary Analysis of English Podcasts

Ulugbek Nurmukhamedov
Northeastern Illinois University, USA

Shoaziz Sharakhimov
Tashkent Medical Academy, Uzbekistan

Abstract

In addition to movies, television programs, and TED Talks presentations, podcasts are an increasingly popular form of media that promotes authentic public discourse for diverse audiences, including university professors and students. However, English language teachers in the English as a second language/English as a foreign language contexts might wonder: "Howdo I know that my students can handle the vocabulary demands of podcasts?" To answer that question, we have analyzed a 1,137,163-word corpus comprising transcripts from 170 podcast episodes derived from the following popular podcasts: *Freakonomics*; *Fresh Air*; *Invisibilia*; *Hidden Brain*; *How I Built This*; *Radiolab*; *TED Radio Hour*; *This American Life*; and *Today Explained*. The results showed that knowledge about the most frequent 3000 word families plus proper nouns (PN), marginal words (MW), transparent compounds (TC), and acronyms (AC) provided 96.75%coverage, and knowledge about the most frequent 5000 word families, including PN, MW, TC, and AC provided 98.26% coverage. The analysis also showed that there is some variation in coverage among podcast types. The pedagogical implications for teaching and learning vocabularyvia podcasts are discussed.

Keywords

Lexical coverage, vocabulary profiling, podcasts, vocabulary instruction

homepage: https://ulugbeknur.wixsite.com/profile

Corresponding author:
Ulugbek Nurmukhamedov, Northeastern Illinois University, 5500 North St. Louis Avenue, Chicago, IL 60625-4699, USA.
Email: u-nurmukhamedov@neiu.edu

Introduction

There is a growing number of language resources on the Web. One of them is a podcast that is gaining popularity among teachers as well as language learners (see Steel and Levy, 2013). *A podcast* can be defined as "a multimedia file, such as a radio program or video, that can be downloaded or streamed from the Internet onto a computer or mobile device" (see *Macmillan Dictionary*). As of 2020, there were around 850,000 active podcasts and around 30 million episodes (see infographic at Music Oomph). Research suggests that 51% of the United States population listened to a podcast or is familiar with the term *podcasting* (see Edison Research Project, 2019). Podcasts are increasingly gaining popularity thanks to their format and the way they deliver content to their listeners. Podcast users do not have to sit and listen to a podcast at a specified time; instead, they can listen to their preferred podcasts "while riding the bus or subway, walking across campus or through a shopping mall" (Thorne and Payne, 2005: 386). Since many professional groups (e.g., TED Talks and British Council), entities (e.g., National Public Radio), periodicals (e.g., *The Economist* and *The New Yorker*) and universities (e.g., Harvard and Massachusetts Institute of Technology) offer their podcasts for free, listeners have access to a wealth of information on topics including society, culture, business, comedy, politics, and health.

Although podcasts are freely available and meant to provide authentic aural input, comprehending the information in podcast episodes can be challenging for second language (L2) learners. One reason is that very few podcasts are created with language learners in mind while there are many general-audience podcasts that are accessible to native speakers and/or expert users of English (see Nurmukhamedov and Sadler, 2011). Another reason is related to whether learners have sufficient vocabulary size (numbers of words that learners know) to listen to general-audience podcasts. It is essential for teachers to identify the numbers of words learners need to know to comprehend a podcast episode before teachers assign a podcast. The aim of this study is to determine the number of words needed to understand general-audience podcasts and offer suggestions on how teachers can prepare learners for the lexical demands of podcasts in English.

How Many Words Do Learners Need to Know to Understand Audio Passages?

Previous research found a linear relationship between vocabulary and listening comprehension (Stæhr, 2009; van Zeeland and Schmitt, 2013). There are many ways to investigate this relationship, one of them being *lexical coverage*, which can be done by calculating the percentage of known words in a text. Lexical coverage is useful because "it indicates the vocabulary size necessary for comprehension of a text" (Rodgers and Webb, 2016: 165). Studies of lexical coverage have consistently used 95% and 98% coverage figures in determining the number of words learners need to know in listening passages because these coverage figures are necessary for "adequate" comprehension. Generally, 95% coverage means that one word out of 20 will be unknown, while in 98% coverage, one word out of 50 will be unknown (Nation and Webb, 2011). It is important to note that "adequate comprehension" is not clearly defined in the literature and may

range in levels from *reasonable*, *good* to *high* listening comprehension (see Nurmukhamedov and Webb, 2019). In a study by van Zeeland and Schmitt (2013), English as a second language (ESL) participants ($n = 40$) listened to four anecdotes and took a ten-item multiple choice listening comprehension test. The researchers replaced some words with non-words in the anecdotes. Thus, each anecdote's lexical coverage varied: 100, 98, 95, and 90. The idea behind this manipulation was to find out whether knowing more words in the anecdotes would lead to better comprehension (i.e., more correct answers in the comprehension test). van Zeeland and Schmitt (2013) found that adequate listening comprehension is achievable when lexical coverage levels are at 95% and 98%. This is supported by Stæhr (2009), who found that 98% coverage provided high-level listening comprehension. Even when the coverage level is less than 95%, learners can still achieve adequate comprehension of spoken input (Bonk, 2000); however, learners need to have a repertoire of coping strategies in different listening scenarios. In sum, it can be concluded that 95% coverage is sufficient for good listening comprehension while 98% coverage is necessary for high-level listening comprehension of informal narratives.

By drawing on the findings of vocabulary knowledge and L2 listening comprehension, several researchers have explored the vocabulary size needed for general spoken English, academic lectures, television (TV) programs, and movies in English. Webb and Rodgers (2009a) examined the vocabulary size necessary to comprehend movie dialogues in English. This study was driven by a popularly held belief that watching more movies will improve a learner's listening, thus eventually improving a learner's vocabulary. They collected 314 movies from 13 different genres (e.g., animation and horror) and compiled a corpus comprising 2,841,887 running words. The results showed that for 95% coverage, a movie viewer needs to know 3000 word families, including proper nouns (PN) and marginal words (MW). For 98% coverage, a movie viewer needs to know 6000 word families, including PN and MW. It should be noted that word familiesare a typical unit of counting in lexical coverage studies. A word family is made up of aheadword, its inflections as well as its derivations (e.g., *visible*, *visibility*, *visibly*, *invisi-ble*, and *invisibility*). It is assumed that if a learner knows one word of a word family, it implies that s/he recognizes all other members of that family (Bauer and Nation, 1993). In addition, PN, MW/interjections, and swear words are generally analyzed and reportedin lexical coverage studies.

In a subsequent study, Webb and Rodgers (2009b) identified the vocabulary coverage necessary for TV programs in English. Their findings revealed that 3000 word families plus PN and MW were necessary to reach 95% coverage of 88 TV programs, and 7000 word families plus PN and MW were necessary to reach 98% coverage. In addition to movies and TV programs, researchers were generally interested in lexical coverage of other listening/spoken discourse types (see Table 1).

Table 1 provides useful information about the methodological practices and outcomes of lexical coverage studies that pertain to aural discourse types. First, the number of words necessary to reach 95% and 98% coverage points varies from one genre to another (see Table 1). For example, compared with movies and TV programs in English, learners are encouraged to have a larger vocabulary size to listen to spoken academic lectures typically used in university-based settings (Dang and Webb, 2014) and TED

Table 1. Lexical coverage of a spoken discourse.

Previous literature (alphabetical order)	Lexical coverage topics	Coverage figures (%) and number of word families
Adolphs and Schmitt (2003)	General spoken English	(95) 3000*/ (96) 5000*
Al-Surmi (2014)	Soap opera	(95.49) 2000/ (98.19) 5000
	Sitcom	(95.06) 2000/ (98.07) 7000
Dang and Webb (2014)	Academic spoken English	(96.05) 4000/ (98.00) 8000
Nation (2006)	General spoken English	(98) 6000–7000
Nurmukhamedov (2017)	TED Talks Presentations	(95.89) 4000/ (98.07) 8000
Webb and Paribakht (2015)	Listening passages in CanTEST (an English proficiency test used for university admission purposes in Canada)	(95.39) 4000/ (98.04) 10,000
Webb and Rodgers (2009a)	Movies in English	(95.76) 3000/ (98.15) 6000
Webb and Rodgers (2009b)	Television programs in English	(95.45) 3000/ (98.27) 7000

Note: * proper nouns and marginal words are not included.

Talks presentations (Nurmukhamedov, 2017). The current study aims to contribute to the available empirical body of knowledge about lexical coverage by determining the number of words necessary to understand general-audience podcasts in English.

Podcast-Based Learning and its Advantages

Compared to other technological tools, podcasts have a number of advantages. First, podcasts are portable, thus learners can carry hundreds and thousands of hours of listening input—ranging from vocabulary or grammar points to conversations on various topics—in their mobile devices. Podcasts enable students to listen to episodes anywhere at any time. Second, podcasts are nearly always free, which makes them an economic "software" for learning. The Internet is certainly needed to download podcasts or newly released episodes; once learners subscribe to their favorite podcasts, valuable and rich content come to their computer or mobile devices on a weekly/daily basis. Third, podcasts are popular among university students. In a survey study that examined the use of technologies by learners ($n = 587$) of a range of languages at a major Australian university between 2006 and 2011, Steel and Levy (2013) found that around half of the participants (238 out of 587) indicated that they used podcasts to learn foreign languages. Others suggest that podcasts positively affect students' study habits because podcasts enable learners to listen to episodes anywhere at any time, multitask (listen and take notes), listen repeatedly (some portion of an episode) as well as control the speed, if necessary, for better comprehension (Fernandez et al., 2015).

Fourth, podcasts offer a wealth of authentic recordings of natural speech. In fact, "interviews, phone calls, and other kinds of social exchanges are common features of podcasts and can serve as models for the L2 listener" (McBride, 2009: 158). Several researchers (Alm, 2013; Yeh, 2013) reported the potential contribution of podcasts to extensive listening, which is defined as "doing a lot of easy, comprehensible, and enjoyable listening practice" (Chang and Millet, 2014: 31). For example, Yeh (2013) examined

why intermediate-level undergraduate students (n = 23) in a Taiwanese university use

podcasts for listening practice. She also observed the students' attitudes towards podcasting as an educational resource. She found that listening to podcasts outside of regular class sessions provided the students with opportunities to listen to authentic speech, engage in topics of personal/academic interests, and carry out meaningful practices (e.g., use a dictionary for the meanings of new words; take notes, etc.) while or after listening to a podcast. In a study that involved 28 intermediate learners of German, Alm (2013) investigated the students' out-of-class listening activities when they self-select their own podcasts. In her study, the participants used a personal blog about their podcast use, commented on each other's blog posts, wrote a review for the podcasts they listened to, and completed a survey about their overall experiences about their podcast-based listening activities. Among many interesting findings, Alm (2013) found that her participants preferred choosing their own podcasts because the self-selected podcasts enabled them to align the podcast episodes with their individual listening goals and personal learning practices. Due to their digitized nature, podcasts allow learners to focus on words/phrases by listening to an isolated passage/segment several times, slowing down speech rate, and pausing between segments to process the information in podcast episodes.

Extensive listening to podcasts may create opportunities for incidental vocabulary learning to occur (Meier, 2015). While listening to large quantities of aural target language input through podcasts, learners might encounter some specialized and/or lower frequency words which they have not heard or learned before. The idea behind the incidental vocabulary learning is that learners are more likely to acquire new words without conscious attention to vocabulary when the words are repeatedly encountered in context (Nation and Webb, 2011). If podcasts are to be used as extensive listening material to improve learners' vocabulary, teachers might want to know the vocabulary demands of podcasts before assigning podcasts to their students. No study has so far examined the lexical coverage of general-audience podcasts in English. The present study will address this gap by answering the following research questions:

(1) How many words do English language learners need to know to understand English podcasts?
(2) Will different podcasts draw on different vocabulary sizes to reach 95% and 98% coverage?

Methodology

Materials

Table 2 includes a list of podcasts selected for the present study based on the following factors: popularity; availability of transcripts; and a wide range of topics. As of May 2020, all of the selected podcasts for the study were in the top 100 podcast shows list according to a web-based radio service platform called Stitcher (see https://tinyurl.com/top100podcasts). Furthermore, except *Today Explained*, all of the podcasts listed in Table 2 were in the top 100 podcast shows in the iTunes charts (as of May 2020). Table 2 gives general information about the selected podcasts for the study.

Table 2. General information about podcast corpus.

Podcasts	Number of episodes	Number of words	Number of hours[a]
Freakonomics	20	146,500	15:29:32
Fresh Air	20	131,057	11:53:28
Hidden Brain	20	98,133	09:27:43
How I Built This	20	161,059	15:13:28
Invisibilia	20	136,477	13:51:34
Radiolab	10	14,137	01:12:58
TED Radio Hour	20	163,256	16:28:51
This American Life	20	212,020	20:37:08
Today Explained	20	74,524	08:17:53
Total	170	1,137,163	112:33:05

Note: [a]format of time units: hours; minutes; and seconds.

The selected podcasts for the current study were in the category of general-audience podcasts, according to the podcast category classifications by Nurmukhamedov and Sadler (2011). Unlike language learning/teaching podcasts specifically prepared by language teachers with language learners in mind, the selected podcasts for the current podcast corpus are primarily designed with native speakers in mind and cover a range of topics ranging from politics and finance to culture and psychology. The Online Appendix contains brief information about the selected podcasts and their foci. Podcast episodes from nine different podcasts were used to ensure that the podcast corpus is representative of the general-audience podcast domain.

To create the podcast corpus, the transcripts of 170 podcasts were downloaded and analyzed in this study. For convenience purposes, the podcast episodes that provided free full transcripts were selected for the podcast corpus. Twenty episodes for each podcast were used except *Radiolab*, which contained transcripts for 10 episodes. In total, the podcast corpus contained 1,137,163 words. Table 2 shows that the selected podcast episodes had a total running time of 112 hours and 33 minutes and an average running time of 40 minutes. The podcast episodes used to compile the podcast corpus can be viewed at https://tinyurl.com/podtranscripts/. The link will take a reader to the first author's Google Drive folder which contains detailed information about each podcast episode (e.g., episode title, duration, etc.).

Analysis

The AntWordProfiler (Anthony, 2014) was used to analyze the podcast transcripts. This is a computer program that lists all of the words in a text according to their frequency and how many times they were used in word lists. The BNC/COCA word family lists, that is, twenty-five 1000-words-frequency lists, were used with the AntWordProfiler software to show the 1000-words level (1000–25,000) at which the words in the podcast transcripts occurred. Nation (2018) created the twenty-five 1000-words lists based on

frequency and range of occurrence of words in the British National Corpus (BNC), and Corpus of Contemporary of American English, commonly known as the COCA corpus (Davies, 2008–2020). The first two 1000 word family lists in the BNC/COCA list were sourced from written, and spoken discourse such as movies, TV programs, and face-to-face/telephone conversations at these levels. The primary goal of creating the first two 1000-words-family list alternative to the so called "General Service List" (West, 1953) was to reflect the balanced word lists that represent "a set of high frequency word lists that were suitable for teaching English as a foreign language and language course design" (see Nation, 2016: 134).

The four additional lists in the BNC/COCA word family lists contain PN (e.g., Clinton and Donovan), MW (e.g., *aha*, *oh*, and *shh-shh*), transparent compounds (TC) (e.g., bird-house and goalkeeper), and acronyms (AC) (e.g., GDP and NHBC). One additional list—*Not in the Lists*—contains either less frequent words or specialized vocabulary (e.g., *artsy*, *bazillion*, *oxycontin*, *vaping*, etc.). The BNC/COCA word lists can be downloaded from Paul Nation's website: https://www.victoria.ac.nz/lals/about/staff/paul-nation

In order to provide a reliable assessment of the vocabulary in podcasts, a number of modifications had to be made in the podcast transcripts. First, contractions (e.g., she's and could've), connected speech (e.g., gonna and kinda), and hyphenated words (e.g., one-bedroom apartment and sight-impaired people) were changed (e.g., she is, could have, going to, kind of, one-bedroom apartment, and sight impaired people) to conform with spelling used in the BNC/COCA word lists. If the spellings of the contractions, connected speech, and hyphenated words were not changed, the words would have been classified as *Not in the List* words (also known as less frequent than the most frequent 25000 word families). Second, PN and AC used in the podcast corpus but not classified as PN (e.g., Zuckerberg, Kushner, etc.) and abbreviations (e.g., COVID, DACA, etc.) according to the BNC/COCA word lists were manually identified and were reclassified as PN and AC, and then added to the ever-growing list of PN and AC totals. In addition, company names (e.g., Uber, U-Haul, etc.), social networking services (e.g., Instagram, Facebook, etc.), ethnic names (e.g., Erdogan, Huma, etc.), and destinations (e.g., Charlottesville, Busoler, etc.) were reclassified as PN and added to the PN' totals. Otherwise, a number of PN would have been classified by AntWordProfiler as *Not in the Lists*: words either not found in the BNC/COCA word lists and/or less frequent than the most frequent 25000 word families.

Results and Discussion

In answer to the first research question, the data in Table 3 indicate that with a vocabulary of 3000 word families and PN, MW, TC, and AC, 96.75% of the words would be known in the podcast corpus. With a vocabulary of 5000 word families and PN, MW, TC, and AC, 98.26% of the words would be known.

Table 3 shows the cumulative coverage for the tokens in the podcasts with and without the PN, MW, TC, and AC. As Table 3 indicates, the first 1000 word families plus PN, MW, TC, and AC accounted for 88.38% of the words, the second 1000 word families plus PN, MW, TC, and AC accounted for 93.76% of the words, and the third 1000 word families plus PN, MW, TC, and AC accounted for 96.75% of the words. Table 3 also

Table 3. Cumulative coverage in percentages of all podcast episodes, with and without proper nouns (PN), marginal words (MW), transparent compounds (TC), and acronyms (AC).

Word list	Coverage without PN, MW, TC, and AC	Coverage including PN, MW, TC, and AC
1000	82.80	88.38
2000	88.18	93.76
3000	91.17	96.75[a]
4000	92.15	97.73
5000	92.68	98.26[b]
6000	93.07	98.65
7000	93.30	98.88
8000	93.48	99.06
9000	93.60	99.18
10,000	93.71	99.29
11,000	93.79	99.37
12,000	93.86	99.44
13,000	93.92	99.50
14,000	93.96	99.54
15,000	93.98	99.56
16,000	94.00	99.58
17,000	94.02	99.60
18,000	94.12	99.70
19,000	94.13	99.71
20,000	94.14	99.72
21,000	94.15	99.73
22,000	94.16	99.74
23,000	94.17	99.74
24,000	94.17	99.74
25,000	94.17	99.74
PN	4.82	
MW	0.24	
TC	0.33	
AC	0.19	
Not in the lists	0.26	100

Note: [a]reaching 95% coverage; and [b]reaching 98% coverage.

indicates that with a vocabulary of 5000 word families and PN, MW, TC, and AC, 98.26% of the words would be known. This means that learners will need knowledge about 2000 more word families to reach from 95% to 98% coverage. Although the difference is only 3% (from 95% to 98% coverage points), it should be pointed out that the most frequent 3000 word families cover a large proportion of a text while a large number of infrequent word families covers a small proportion of the text (Nation and Webb, 2011). In sum, learners of English would therefore need a vocabulary of the most frequent 3000 and 5000 word families plus the knowledge about PN, MW, TC, and AC to gain 95% and 98% coverage, respectively.

Table 4. Cumulative coverage in percentages of each podcast category, including proper nouns (PN), marginal words (MW), transparent compounds (TC), and acronyms (AC).

Podcasts	1000 PN, MW, TC, AC	2000 PN, MW, TC, AC	3000 PN, MW, TC, AC	4000 PN, MW, TC, AC	5000 PN, MW, TC, AC	6000 PN, MW, TC, AC
Freakonomics	84.15	92.06	96.39[a]	97.55	98.18[b]	98.62
Fresh Air	87.72	93.07	96.56[a]	97.57	98.11[b]	98.66
Hidden Brain	87.03	93.14	96.67[a]	97.86	98.40[b]	98.79
How I Built This	90.98	95.29[a]	97.36	98.08[b]	98.51	98.75
Invisibilia	88.36	93.92	96.43[a]	97.42	97.87	98.18[b]
Radiolab	89.05	94.17	96.41[a]	97.39	97.85	98.18[b]
TED Radio Hour	88.03	93.45	96.74[a]	97.78	98.26[b]	98.59
This American Life	90.41	94.87	96.95[a]	97.83	98.39[b]	98.82
Today Explained	86.83	92.83	96.58[a]	97.63	98.26[b]	98.77

Note: [a]reaching 95% coverage; and [b]reaching 98% coverage.

Overall, these results suggest that podcasts in English may be slightly less demanding than university-based academic lectures (Dang and Webb, 2014) and TED Talks presentations (Nurmukhamedov, 2017). However, the vocabulary knowledge necessary for English podcasts is similar to the vocabulary demands of movies and TV programs in English (see Webb and Rodgers, 2009a, 2009b). Taken together, these results suggest that English as a foreign language/ESL learners should preferably have a good grasp of the most frequent 3000 word families (e.g., words in the 1000, 2000, and 3000 words levels) if they intend to use audio-based media outlets such as movies, TV programs, and podcasts to further improve their English. Because 98% coverage is recommended for high level of listening comprehension (van Zeeland and Schmitt, 2013), learning the most frequent 5000 word families or more would aid learners in handling the vocabulary demands of most podcasts in English.

In response to the second research question, it can be seen from the data in Table 4 that the vocabulary necessary to reach 95% coverage was fairly consistent among all the podcasts except *How I Built This*. Knowledge about the most frequent 2000 word families, and PN, MW, TC, and AC was needed to reach 95.29% for *How I Built This*. Knowledge about the most frequent 3000 word families plus PN, MW, TC, and AC provided the highest coverage of all the other remaining podcasts.

With regard to the 98% coverage, there were some notable differences among the podcasts. The vocabulary size necessary to reach 98% coverage ranged from 4000 to 6000 word families plus PN, MW, TC, and AC. Knowledge about the most frequent 5000 word families plus PN, MW, TC, and AC was sufficient to reach 98% coverage of *Freakonomics* (98.18%), *Fresh Air* (98.11%), *Hidden Brain* (98.40%), *TED Radio Hour* (98.26%), *This American Life* (98.39%), and *Today Explained* (98.26%). However, a vocabulary of the most frequent 6000 word families plus PN, MW, TC, and AC was necessary to reach 98.18% coverage of two podcasts: *Invisibilia*; and *Radiolab*. In contrast, *How I Built This* was the only podcast, requiring 4000 word families plus PN, MW, TC, and AC for 98% coverage. These results indicate that *How I Built This* is the least demanding podcast in terms of lexical coverage

while *Invisibilia* and *Radiolab* are the most demanding podcasts because they require the largest vocabulary sizes to reach 98% coverage.

The podcasts may have differed in terms of their coverage percentages for the following reasons. First, podcasts differ in terms of their formats. For example, most of the selected podcasts have one host and in each episode the host invites one guest to discuss an issue while podcasts such as *Radiolab* and *Invisibilia* typically have two or more hosts and their episodes feature multiple guests. Second, each podcast focuses on different topics in their episodes (see the Online Appendix for more information). For example, a host of the *How I Built This* podcast invites a founder of a famous company to the show to discuss how his/her company became successful. Two people—usually a host and a company founder(s)—have a conversation about one company's recent history and current condition in each episode. In podcasts such as *Radiolab* and *Invisibilia*, several hosts discuss a wide range of topics (e.g., psychology, sociology, science, and politics) with several guests which may include both expert and non-expert individuals. Third, the variation between the vocabulary sizes necessary to reach 95% and 98% coverage of each podcast supports the idea that "different disciplines may have different lexical demands" (Dang and Webb, 2014: 72). For example, even among movies in English, British movies required more vocabulary (7000 word families) than American movies (6000 word families) to reach 98% coverage; furthermore, *animated* and *war* movies were more lexically demanding than *drama*, *comedy* and *action* (see Webb and Rodgers, 2009a). Similarly, *science-fiction* TV programs in American TV required more vocabulary than *situation comedies* at a 98% coverage level (see Webb and Rodgers, 2009b). Diverse podcast episodes and their content contain a variety of vocabulary, which ultimately affect the lexical coverage analysis.

In addition to the most frequent 3000–5000 word families, variables such as PN, MW, TC, and AC are essential to handle the lexical demands of podcasts for the following reasons. First of all, PN, MW, TC, and AC available in the current study comprise nearly 5.58% of the coverage of the podcast corpus. This percentage amounts to 4882 word types. This seems to be larger than most of the previous studies of lexical coverage that examined movies (3.37%), TV programs (4.04%), TED Talks presentations (0.98%), and academic spoken English (3.33%). The reason is that most of the previous studies have utilized the BNC word lists that contained two additional lists such as PN and MW. The current study used the BNC/COCA word lists, which has two more additional lists: TC and AC. These four additional lists add up to the total coverage. Another reason for the large percentage of PN, MW, TC, and AC in the current study is that the word lists for the PN and AC in the BNC/COCA word lists were expanded before the analysis in order to represent all of the PN and AC used in the podcast episodes. It should be noted that the additional word lists—PN, MW, TC, and AC—had the following distributions per podcast in the current study: *Freakonomics* (3.90%); *Fresh Air* (4.40%); *Hidden Brain* (5.14%); *How I Built This* (5.28%); *Invisibilia* (6.25%); *Radiolab* (9.69%); *TED Radio Hour* (4.56%); *This American Life* (7.70%); and *Today Explained* (6.08%). This expansion would also contribute to more accurate representation of the coverage.

Findings and Pedagogical Implications

The findings in the current study indicate that knowledge about the most frequent 3000 word families and PN, MW, TC, AC provide 95% coverage of general-audience podcasts. Although this coverage is sufficient for adequate comprehension of podcasts, we recommend that teachers aim for a vocabulary size of 5000 word families, which leads to 98% coverage. In their seminal paper, van Zeeland and Schmitt (2013) argue that the learners need to know 98% of the words in a listening passage in order to achieve successful (high) listening comprehension of difficult listening materials. Furthermore, podcasts generally do not contain visual clues or subtitles/captions (i.e., unlike *podcasts*, so-called *vodcasts* do contain video elements). Thus, learners depend on their listening skills and vocabulary knowledge to listen to podcasts; while, movies, TV programs and university-based lectures offer visual support by containing scenes where people interact with each other, show animated objects, and illustrate presentation slides. Teachers also need to ensure that their learners know 5000 or more words before they integrate podcasts into their curriculum and help their students view podcasts as an additional source of input to improve their listening skills, consolidate the use of their previously learned words, and expand on their vocabulary size. Here are some suggestions on how our findings can inform language teaching.

First, guiding learners to select a level-appropriate podcast(s) is crucial because general-audience podcasts available on the Web are typically designed for native speakers/expert users of English and have not yet been structured according to their difficulty levels. If podcast episodes are not carefully selected, learners might find it "difficult to catch all of the words while listening to authentic materials" (Meier, 2015: 72); thus, listening to an unsuitable podcast might result in frustrating learning experience (Alm, 2013). We suggest that listening to a less-lexically-demanding podcast (or an episode) may ease the burden of comprehension. For example, *How I Built This* is the least lexically demanding podcast among the selected podcasts (see Table 4), requiring 4000 word families, plus PN, MW, TC, and AC to reach 98% coverage. Some short episodes from *How I Built This* can be used to attune learners' ears to this L2 aural mode (e.g., pronunciation of various speakers, speed in a dialogue, etc.) and familiarize learners with the format, topics, and common story narratives in a typical podcast episode. Once learners practice listening to podcast episodes that require less vocabulary, teachers can move onto podcasts such as *Fresh Air, Hidden Brain*, and/or *TED Radio Hour* that are more lexically demanding than the *How I Built This* podcast.

Second, pre-teaching key or low-frequency vocabulary essential for the comprehension of a selected podcast episode will also ease the burden of comprehension (van Zeeland, 2018; Webb, 2010). Key vocabulary includes words that are essential for the comprehension of a selected episode. To illustrate, a passage on COVID-19 typically contains the following words: *coronavirus*; *emergency*; *lockdown*; *pandemic*; *self-quarantine*; *social distancing*; *vaccine*, etc. Teachers can find low-frequency or academic words in the podcast episode by using Vocabprofiler software which can be found in a free accessible website Compleat Lexical Tutor at https://www.lextutor.ca/vp/eng/ (see Cobb, 2007 for more information). Upon entering a podcast's transcript into Vocabprofiler, this online software will create the frequency profile of the transcript and illustrate a list

of the most frequent 1000 and 2000 words, including the words from Coxhead's (2000) Academic Word List and infrequent words (also known as off list words). The transcripts of the podcast episodes used in the present study can be found at the first author's Google Drive at: https://tinyurl.com/podtranscripts/.

Another approach is to train learners how to use podcast transcripts to support listening to podcasts. During the first week, the teacher selects a podcast episode for a unit and provides learners with the same podcast's written transcript. Learners are asked to follow along with the podcast transcript while listening to the podcast episode. After using this technique with new podcast episodes for a couple of times, the teacher can gradually remove this support (written transcripts) as learners become more familiar with the speech rate, pauses, repetitions, and topic digressions in podcasts. Having students follow transcripts while listening to the podcast may ease the burden of comprehension too. In addition, reading a podcast transcript, which enables learners to see the written forms of words while listening to the same podcast episode, will help them listen to the spoken forms of the words.

Fourth, developing learners' schema before assigning a podcast episode may also help learners turn their attention to vocabulary when they listen to a podcast. Before assigning a podcast episode, teachers may want to provide learners with some key information about the selected episode such as a main idea in the episode, the speaker's background, and credibility about the topic and two or three interesting points from the podcast episode. In addition, teachers–especially in EFL contexts—need to spend some time to introduce PN (Erten and Razi, 2009) and AC for a selected podcast episode to ease students' comprehension of the information presented in the episode. Most podcast episodes include the names of investors, authors, and chief executive officers as well as their inventions. This indicates that not all learners might be familiar with the accomplishments and/or characteristics of individuals such as *Agatha Christie*, *Malala Yousafzai*, and *Mark Zuckerberg*; and most importantly, some learners might not know what some AC (e.g., DACA, LGBT, etc.) stand for. Furthermore, some company and website names (e.g., Instagram, Zappos.com, etc.) should also be introduced if a podcast episode contains company/website names. Because PN carry specific contextual as well as cultural information and the knowledge about PN leads to better comprehension of information in a passage, they need to be explicitly taught to learners (Brown, 2010).

Finally, learners should be notified about the presence of exclamations/interjections (e.g., wow, uh, oh, and mmm) and contractions (e.g., 'm, gonna, and wanna) in podcast episodes. Language learners may feel that spoken English in the podcast episodes is fast, oftentimes hard to understand and learners might complain that "They can't hear each word clearly; instead, the words sound like one long, confusing stream of sound. This is because when people talk normally, their words naturally blend and change in predictable ways" (Yoshida, 2016: 111). Teachers may want to introduce the notion of *connected speech* to learners so that learners notice that in natural/spoken English, native and/or expert users of English blend/contract words (e.g., *that will* becomes *that'll*), link them (e.g., a dog is black and white), and change sounds in words (e.g., *have to* or *used to* become 'hafta' or 'usta') (see Yoshida, 2016 for more information). Listening to a variety of podcasts helps to enable English language learners to "hear" how authentic (informal) speech is supposed to sound (Godsey, 2016). In addition to the explicit instruction in

pronunciation training, especially in the area of *connected speech*, teachers can recommend that their students use a "slow down" feature in their smart phones to "catch" instances of connected speech. This feature—which is nowadays available in many smart phones—helps to enable learners to pause, skip, slow down or speed up a conver-sation as they listen. Such exercises can help learners understand the nature of spoken language and aid in their listening comprehension.

Limitations and Recommendations

Following the methodology developed by the previous lexical coverage studies of spoken language such as university lectures (Dang and Webb, 2014), TED Talks presentations (Nurmukhamedov, 2017), and movies as well as TV programs in English (Webb and Rodgers, 2009a, 2009b), the present corpus-based study examined the number of words necessary to understand 95%–98% of the words in general-audience podcasts. The current study does not claim that the knowledge about the 98% of the words in a podcast would lead to 98% listening comprehension. Learners might understand all the words in a listening passage but not necessarily understand the intended message in the passage (Graham, 2000). Another reason is that vocabulary is one of many factors that affect L2 listening comprehension, among other factors such as learners' overall lan- guage proficiency and metacognitive awareness (i.e., learners' awareness of their abilityto solve a task, strategies employed while listening, etc.) (see Wang and Treffers-Daller,2017). Thus, further research needs to examine more closely the links between a learn- er's actual vocabulary size and his/her listening comprehension of a podcast episode, following the methodology outlined in the works of Stæhr (2009) and van Zeeland and Schmitt (2013).

Acknowledgements

The author(s) thank Teddy Bofman (at Northeastern Illinois University) for her constructive feedback on the first draft of the manuscript, the journal editor, and the anonymous reviewer for his/her constructive suggestions.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by the Summer 2018 Research Stipend granted by Northeastern Illinois University.

ORCID iD

Ulugbek Nurmukhamedov https://orcid.org/0000-0001-9293-2865

Supplemental material

Supplemental material for this article is available online.

References

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

Alm A (2013) Extensive listening 2.0 with foreign language podcasts. *Innovation in Language Leaning and Teaching* 7(3): 266–280.

Al-Surmi M (2014) TV shows, word coverage, and incidental vocabulary learning. In: Bailey K, Damerow R (eds) *Teaching and Learning English in the Arabic-Speaking World*. London: Routledge, 132–147.

Anthony L (2014) AndWordProfiler. (Version 1.4.1) [Computer Software]. Tokyo: Waseda University. Available at: https://www.laurenceanthony.net/software/antwordprofiler/ (accessed August 2020).

Bauer L, Nation ISP (1993) Word families. *International Journal of Lexicography* 6(4): 253–279.

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

Brown D (2010) An improper assumption? The treatment of proper nouns in text coverage counts. *Reading in a Foreign Language* 22(2): 355–361.

Chang A, Millett S (2014) The effect of extensive listening on developing L2 listening fluency: Some hard evidence. *ELT Journal* 68(1): 31–40.

Cobb T (2007) Computing the vocabulary demands of L2 reading. *Language, Learning and Technology* 11(3): 38–63.

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

Dang T, Webb S (2014) The lexical profile of academic spoken English. *English for Academic Purposes* 33: 66–76.

Davies M (2008–2020) The Corpus of Contemporary American English: 560 Million Words, 1990–Present. Available at: http://corpus.byu.edu/coca/ (accessed March 2020).

Edison Research Report (2019) The Spoken Word Audio Report. Available at: https://www.nationalpublicmedia.com/spoken-word-audio-report/ (accessed August 2020).

Erten I, Razi S (2009) The effects of cultural familiarity on reading comprehension. *Reading in a Foreign Language* 21: 60–77.

Fernandez V, Sallan J, and Simo P (2015) Past, present, and future of podcasting in higher education. In: Li M, Zhao Y (eds) *Exploring Learning and Teaching in Higher Education*. Berlin: Springer, 305–330.

Godsey M (2016) The value of using podcasts in class. *The Atlantic*, 17 March. Available at: https://www.theatlantic.com/education/archive/2016/03/the-benefits-of-podcasts-in-class/473925/ (accessed November 2019).

Graham S (2006) Listening comprehension: The learners' perspective. *System* 34(2): 165–182.

McBride K (2009) Podcasts and second language learning: Promoting listening comprehension and intercultural competence. In: Abraham L, Williams L (eds) *Electronic Discourse in Language Learning and Language Teaching*. Amsterdam: John Benjamins, 153–168.

Meier A (2015) L2 incidental vocabulary acquisition through extensive listening to podcasts. *Studies in Applied Linguistics & TESOL* 15: 72–84.

Nation ISP (2016) *Making and Using Word Lists for Language Learning and Testing*. Amsterdam. John Benjamins.

Nation ISP (2018) The BNC/COCA Word Family Lists. Available at: http://www.victoria.ac.nz/lals/about/staff/paul-nation (accessed August 2020).

Nation P, Webb S (2011) *Researching and Analyzing Vocabulary*. Boston, MA: Heinle Cengage Learning.

Nurmukhamedov U (2017) Lexical coverage of TED Talks: Implications for vocabulary instruction. *TESOL Journal* 8(4): 268–290.

Nurmukhamedov U, Sadler R (2011) Podcasts in four categories: Applications to language learning. In: Abdous M, Facer BR (eds) *Academic Podcasting and Mobile Assisted Language Learning: Applications and Outcomes*. Hershey, PA: IGI Global, 176–195.

Nurmukhamedov U, Webb S (2019) Research timeline: Lexical coverage and profiling. *Language Teaching* 52(2): 1–13.

Rodgers MPH, Webb S (2016) Listening to lectures. In: Hyland K, Shaw P (eds) *The Routledge Handbook of English for Academic Purposes*. London: Routledge, 165-176.

Stæhr LS (2009) Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition* 31(4): 577–607.

Steel C, Levy M (2013) Language students and their technologies: Charting the evolution 2006–2011. *ReCALL* 25(3): 306–320.

Thorne S, Payne J (2005) Evolutionary trajectories, internet-mediated expression, and language education. *CALICO Journal* 22(3): 371–397.

van Zeeland H (2018) Vocabulary in listening. In: Renandya W, Hu G (eds) *The TESOL Encyclopedia of English Language Teaching*. London: Wiley: 1–6.

van Zeeland H, Schmitt N (2013) Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics* 3494: 457–479.

Wang Y, Treffers-Daller J (2017) Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System* 65: 139–150.

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

Webb S, Paribakht TS (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 MPH (2009a) The lexical coverage of movies. *Applied Linguistics* 30(3): 407–427.

Webb S, Rodgers MPH (2009b) Vocabulary demands of television programs. *Language Learning* 59(2): 235–366.

West M (1953) *A General Service List of English Words*. London: Longmans, Green & Co.

Whitner G (2020) The meteoric rise of podcasting: Insights about the most compelling audio format. *Music Oomph*, 3 May. Available at: https://musicoomph.com/podcast-statistics/ (accessed August 2020).

Yeh C (2013) An investigation of a podcast learning project for extensive listening. *Language Education in Asia* 4(2): 135–149.

Yoshida M (2016) *Beyond Repeat After Me: Teaching Pronunciation to English Learners*. Alexandria, VA: TESOL Press.

Author biographies

Ulugbek Nurmukhamedov teaches in the MA Teaching English to Speakers of Other Languages program at Northeastern Illinois University, Chicago, USA. His research interests include vocabulary studies and computer-assisted language learning. His articles have been published in *Language Teaching* and *International Journal of Lexicography*. His book (with Randall Sadler) is *New Ways in Teaching with Games* (2020, TESOL Press).

Shoaziz Sharakhimov obtained his MA Teaching English degree from UzSWLU. Currently, he teaches English for Specific Purposes/English for Academic Purposes courses at Tashkent Medical Academy (Tashkent, Uzbekistan). His research interests include second language (L2) assessment and L2 acquisition.