

Investigating the Lexical Demands of English-as-an-Additional-Language and General-Audience Podcasts and Their Potential for Incidental Vocabulary Learning

Masoud Motamedynia
University of Tehran

Naseh Nasrollahi Shahri
San Diego State University

Abstract

This study investigated the lexical demands of English-as-an-additional-language (EAL) and general-audience podcasts and their potential for incidental vocabulary learning. Two corpora (i.e., one EAL and one general-audience) comprising 1,188,512 tokens were analyzed to determine the necessary vocabulary knowledge to reach 90% and 95% coverage. The results indicated that 2,000 and 3,000 word families plus proper nouns (PN), marginal words (MW), transparent compounds (TC), and acronyms (AC) covered 90% and 95% of words in podcasts, respectively. The results also showed that EAL and general-audience podcasts required 1,000 and 2,000 words families to reach 90% coverage, respectively. Regarding 95% coverage, knowledge of 2,000 (EAL) and 3,000 (general-audience) word families was required. The results also demonstrated almost 60% of word families from the 2,000-word level were encountered 15+ times in each corpus, suggesting podcasts may hold relatively great potential for learning such words incidentally. Furthermore, our findings indicated that there was some potential for incidentally learning words from the 3,000-word level in both corpora, while general-audience podcasts may hold greater potential in this regard. Implications for using podcasts in language pedagogy are also discussed.

Résumé

Cette étude a examiné les niveaux lexicaux des podcasts l'anglais comme langue additionnelle (ALA) et des podcasts audience-générale et leurs potentiels pour l'apprentissage incident du vocabulaire. Deux corpus (un corpus ALA et un corpus audience-générale) qui ont consisté 1 188 512 mots ont été analysés pour déterminer la connaissance du vocabulaire nécessaire pour atteindre une couverture de 90 % et 95 %. Les résultats ont révélé que 2 000 familles de mots et 3 000 familles de mots avec noms propres, interjections, composés transparents et acronymes couvraient respectivement 90 % et 95 % des mots des podcasts. De plus, les résultats ont indiqué que les podcasts ALA et les podcasts audience-générale demandaient respectivement 1 000 familles de mots et 2 000 familles de mots pour atteindre une couverture de 90 %. Pour une couverture de 95 %, la connaissance de 2 000 familles de mots (ALA) et de 3 000 familles de mots (audience-générale) était requise. De plus, les résultats ont démontré que près de 60 % de 2 000 familles de mots les plus fréquentes ont été rencontrées plus de 15 fois dans chaque corpus. Ces résultats indiquent que l'écoute de podcasts peut avoir un potentiel relativement fort d'apprentissage incident de tels mots. De plus, les résultats ont indiqué qu'il y avait un potentiel d'apprentissage incident de 3 000 familles de mots les plus fréquentes dans chaque corpus, tandis que les podcasts

audience-générale pourraient avoir un plus grand potentiel à cet égard. Les implications de l'utilisation des podcasts dans la pédagogie des langues sont également discutées.

Investigating the Lexical Demands of English-as-an-Additional-Language and General-Audience Podcasts and Their Potential for Incidental Vocabulary Learning

Introduction

Advancements in information technology and easy access to the Internet have made available vast quantities of information. The rapid growth of technology has provided learners with ample opportunity to learn a language outside the classroom and at their desired pace. Podcasts are one among many technological tools that are being given more attention in additional language learning and teaching as they are an authentic source of linguistic input. Macmillan Dictionary (see <https://macmillandictionary.com>) defines a podcast as “a multimedia file, such as a radio programme or video, that can be downloaded or streamed from the internet onto a computer or mobile device”. Podcasting has been growing rapidly globally, with more than 2,000,000 podcasts (over 48 million episodes) available as of April 2021 (Winn, 2021).

Nurmukhamedov and Sadler (2011) proposed a taxonomy of podcasts with four categories: a) discrete; b) ESL-focused; c) ESL-super; and d) general-audience. The first category contains podcasts usually with a narrow focus produced particularly for expert users of English and characterized by less frequent words, thus probably useful for learners with specific purposes. In contrast, what they referred to as ESL-focused and ESL-super are podcasts both created specifically for language learning and teaching purposes, offering linguistic input often modified to reduce barriers to comprehension. The two categories, however, differ in that the latter provides learners with many different free podcast options, and may or may not be accompanied by language-focused exercises. Although Nurmukhamedov and Sadler (2011) consider these two types of podcasts to be slightly different, we put them under the umbrella term “English-as-an-additional language (EAL) podcasts” as both are produced for language learning and teaching purposes. The reason for this labelling is that “EAL” can be a more strength-based term because it encompasses a more diverse profile of listeners across contexts such as Canada and moves away from deficit views of multilingualism. In contrast, general-audience podcasts are those created without any language learning and teaching purposes. Therefore, they provide unmodified input regarding vocabulary complexity and speech rate, and as the name suggests, their potential listeners can include a wide range of people. All types can provide learners with authentic auditory linguistic input; however, learners’ proficiency levels should be given attention when using each category (Nurmukhamedov & Sadler, 2011).

While it is undoubtedly true that there are manifold factors influencing comprehension, such as prior knowledge (Cervetti & Wright, 2020), vocabulary knowledge can plausibly be the most significant factor (Laufer & Sim, 1985). Research has highlighted the significance of L2 lexical knowledge as an important predictor of success in reading (e.g., Qian, 2002) and listening comprehension (e.g., Matthews, 2018) as well as in oral and written production (Douglas, 2016). Although podcasts are popular among young people, especially university students (Nurmukhamedov & Sharakhimov, 2021), learners of different age groups may find them to be a useful language learning tool. Knowing the lexical demands of podcasts can be an important predictor of their suitability

as a source of linguistic input as it shows the vocabulary size required for their comprehension (Rodgers & Webb, 2016).

Despite the well-established research into the lexical load of various discourse types, only one study (Nurmukhamedov & Sharakhimov, 2021) to date has determined the vocabulary size needed for comprehending podcasts, making it a relatively underexamined area of exploration. Furthermore, no study thus far has explored the difference in the lexical load between EAL and general-audience podcasts, a gap that this study aims to bridge. Drawing a comparison between these two categories of podcasts regarding their lexical demands can be important as they both have a wide audience and are reliable sources of linguistic input. The results of this study will hopefully contribute to the literature on the lexical profiling of various discourse types, and help learners, teachers, and researchers to set vocabulary learning goals upon using podcasts for language learning and teaching purposes. This study also attempts to address another gap in the vocabulary learning literature through a corpus-based investigation of podcasts' potential, both EAL and general-audience, for incidental vocabulary learning.

Literature Review

Lexical Coverage

One line of vocabulary research has addressed lexical coverage to determine the thresholds necessary for reasonable and optimal comprehension of written and spoken sources of linguistic input. Webb (2010, p. 498) defined coverage as “the percentage of known words in a text” and considered it a useful measurement tool as it assists in determining the extent to which learners can comprehend a text and incidentally learn words from it. There have been three seminal studies attempting to explore the effect of lexical coverage on reading comprehension (Hu & Nation, 2000; Laufer, 1989; Schmitt et al., 2011). In one of the earliest studies on lexical coverage, Laufer (1989) found that at 95% text coverage the great majority of the participants scored 55% or higher on a reading comprehension test. In Hu and Nation's (2000) study, none of the participants achieved sufficient comprehension at 80% lexical coverage. Furthermore, only a small number of the participants gained adequate comprehension at 90% coverage, and even at 95% coverage most of the participants did not show enough comprehension. It was at 100% coverage, however, that the majority of the participants gain adequate comprehension. In another study, Schmitt et al. (2011) found that as there was a 1% increase in coverage between 90% and 100% intervals, the participants achieved greater comprehension, suggesting that there is a linear relationship between the two variables.

Thus far, only two studies have addressed the interaction between lexical coverage and listening comprehension. Bonk (2000) found that the participants in his study showed different levels of comprehension at various coverage points. Accordingly, he concluded that “good comprehension seldom occurred with text-lexis familiarity levels lower than 75 percent, but occurred frequently at 90+ percent levels” (p. 14). In another study, van Zeeland and Schmitt (2013b) found that a large number of the participants achieved sufficient comprehension at 90% and 95% coverage levels. The mean scores at 90%, 95%, 98%, and 100% coverage levels were 7.35, 7.65, 8.22, and 9.62, respectively. Their results also showed that although most of the participants, both native and non-native, showed

adequate comprehension at 90% coverage, there was a great deal of variation in non-native participants' test scores at this level. However, there was far less variation in their scores at 95% coverage. Therefore, they concluded that 95% was the most appropriate coverage level in relation to listening comprehension. More recently, Durbahn et al. (2020) conducted a study on the relationship between lexical coverage and viewing comprehension. Their results showed that the participants' comprehension of an English documentary increased from 62% to 92% when there was an increase in the lexical coverage from 87% to 99%. They concluded that less lexical coverage might be required for sufficient viewing comprehension than for reading. They also suggested that the coverage level needed for adequate viewing comprehension was 90%, which was in line with van Zeeland and Schmitt's (2013b) study.

Taken together, research has shown that 95% and 98% coverage levels might be required for reasonable (Laufer, 1989) and ideal, unassisted reading comprehension (Hu & Nation, 2000; Schmitt et al., 2011), respectively. Regarding listening comprehension, however, research has indicated that coverage levels between 90% and 95% might be necessary to gain adequate comprehension (Bonk, 2000; van Zeeland & Schmitt, 2013b). Furthermore, coverage of 90% has been suggested to provide adequate viewing comprehension (Durbahn et al., 2020).

The Lexical Profiles of Different Discourse Types

Much research has been undertaken to find the vocabulary sizes required to reach adequate and ideal coverage levels in various discourse types. As mentioned above, research has suggested different coverage levels for different modes of input: audiovisual, auditory, and written.

Regarding 90% coverage considered adequate for viewing and listening comprehension, the required vocabulary knowledge has been suggested to include 1,000–2,000 word families. Accordingly, teacher-selected songs (Tegge, 2017), teacher talk (Horst, 2010), and IELTS listening (section 1) (Phung & Ha, 2022) were lexically least demanding, requiring 1,000 word families, while the majority of discourse types needed the most frequent 2,000 word families to reach this coverage level: television programs (Webb & Rodgers, 2009b), movies (Webb & Rodgers, 2009a), everyday spoken English (Nation, 2006), pop charts songs (Tegge, 2017), general-audience podcasts (Nurmukhamedov & Sharakhimov, 2021), academic lectures (Dang & Webb, 2014), rap songs (Tegge & Coxhead, 2021), TED Talks (Nurmukhamedov, 2017), CanTEST listening (Webb & Paribakht, 2015), TOEFL iBT listening (Kaneko, 2015), IELTS listening (Sections 2, 3, and 4) (Phung & Ha, 2022), and video games (Rodgers & Heidt, 2020).

Considering 95% coverage, the necessary vocabulary sizes have been suggested to be between 2,000 and 4,000 word families. Teacher-selected songs (Tegge, 2017), teacher talk (Horst, 2010) and IELTS listening (section 1) (Phung & Ha, 2022) required 2,000 word families. On the other hand, most audiovisual and auditory genres needed 3,000 word families to reach the same coverage point: television programs (Webb & Rodgers, 2009b), movies (Webb & Rodgers, 2009a), everyday spoken English (Nation, 2006), pop charts songs (Tegge, 2017), general-audience podcasts (Nurmukhamedov & Sharakhimov, 2021), TOEFL iBT listening (Kaneko, 2015), and IELTS listening (Sections 2, 3, and 4) (Phung & Ha). Moreover, the most demanding genres have been suggested to include academic

lectures (Dang & Webb, 2014), TED Talks (Nurmukhamedov, 2017), rap songs (Tegge & Coxhead, 2021), and CanTEST listening (Webb & Paribakht, 2015) needing 4,000 word families and video games (Rodgers & Heidt, 2020) requiring 5,000 word families.

However, knowledge of 2,000-10,000 word families has been suggested to reach 95% and 98% coverage in written discourse. Accordingly, knowledge of 2,000 word families was required to reach 95% coverage of graded readers (Webb & Macalister, 2013), whereas a number of discourse types needed 3,000 word families to reach the same threshold. These included high-school textbooks (Nguyen, 2020), children and adult literature (Webb & Macalister, 2013), and university student essays (Douglas, 2013). Other more demanding genres included novels and newspapers (Nation, 2006) requiring 4000 word families. Regarding the 98% threshold, 3,000 word families provided 98% coverage of graded readers (Webb & Macalister, 2013). More demanding genres included high-school textbooks (Nguyen, 2020), and university student essays (Douglas, 2013) requiring the most frequent 5,000 word families. The most demanding genres included novels and newspapers (Nation, 2006), requiring 8,000 word families, and children and adult literature (Webb & Macalister, 2013) with 10,000 word families.

Table 1 summarizes the results of the studies on the lexical demands of auditory sources of linguistic input.

Table 1

Vocabulary Knowledge Necessary to Reach 90% and 95% Coverage in Different Auditory Discourse Types

Discourse type	90%	95%	Study
General-Audience Podcasts	2,000+PN/MW/TC/AC	3,000+PN/MW/TC/AC	Nurmukhamedov & Sharakhimov (2021)
Rap Songs	2,000+PN/MW/TC/AC	4,000+PN/MW/TC/AC	Tegge & Coxhead (2021)
Pop Songs	2,000+PN/MW/TC	3,000+PN/MW/TC	Tegge (2017)
Teacher-Selected Songs	1,000+PN/MW/TC	2,000+PN/MW/TC	Tegge (2017)
Academic Lectures	2,000+PN/MW	4,000+PN/MW	Dang & Webb (2014)
Everyday Spoken English	2,000+PN	3,000+PN	Nation (2006)
Teacher Talk	1,000	2,000	Horst (2010)
TOEFL iBT listening	2,000+PN/MW	3,000+PN/MW	Kaneko (2015)
CanTEST listening	2,000+ PN/MW	4,000+PN/MW	Webb & Paribakht (2015)
IELTS listening (four sections)	1,000–2,000+ PN/MW/TC/AC	2,000–3,000+ PN/MW/TC/AC	Phung & Ha (2022)

Note: PN: Proper Nouns; MW: Marginal Words; TC: Transparent Compounds; AC: Acronyms

Word Frequency Lists

Word frequency lists are needed to analyze the vocabulary demand of a text. Nation's (2004) frequency lists are an example of such word lists, classifying words based on frequency and range data gathered from an analysis of the British National Corpus. The practice of utilizing such lists to determine the vocabulary load stems from the notion that the higher the frequency of a word in general language use, the more likely it is to be learned early in the learning process (Nation, 2006). As these lists were developed from a mainly British corpus, they were more representative of British English than American English, a limitation in studies trying to assess the lexical demand of American discourse. Later, Nation (2012, 2017) developed the BNC/COCA word lists based on an analysis of the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA).

Schmitt and Schmitt (2014) proposed the traditional high-/low-frequency dichotomy be reassessed. Considering a number of factors like "dictionary defining vocabulary" and "the amount of vocabulary necessary for English usage" (p. 484), they suggested that high-frequency vocabulary should comprise the most frequent 3,000 word families. However, Dang and Webb (2016) proposed that high-frequency vocabulary should include the most frequent 1,000 word families. They argued so because the coverage provided by the frequency bands beyond the most frequent 1,000 word families drops drastically. More recently, Dang et al. (2020) adopted an innovative, cross-disciplinary approach to exploring what might constitute high-frequency vocabulary. Bringing together data from research on corpus linguistics, and teacher and learner cognition, they suggested that a word list containing the BNC/COCA most frequent 2,000 word families should be considered the more useful list of high-frequency vocabulary.

Incidental Vocabulary Learning from Listening

Incidental vocabulary learning is a line of research that is receiving much attention, defined as "learning words without deliberate decision to commit information to memory" (Laufer & Hulstijn, 2001, p. 11). Research into incidental vocabulary learning through reading is well-established (e.g., Freimuth, 2020; Pellicer-Sánchez & Schmitt, 2010). Recently, a number of studies have also focused on incidental vocabulary learning from listening to different sources of linguistic input (e.g., Feng & Webb, 2019; Horst, 2010; Jin & Webb, 2020; Pavia et al., 2019; Peters & Webb, 2018; Rodgers & Webb, 2020; van Zeeland & Schmitt, 2013a; Vidal, 2011).

Many factors have been suggested to affect incidental vocabulary learning among which frequency of occurrence is considered a significant one. As research findings have greatly differed regarding the precise frequency of occurrence needed for incidental vocabulary learning, it may not be possible to determine the exact number. Chen and Truscott (2010) stated that "the goal of research should not be to identify a definitive number of exposures needed but rather to understand a complex process involving multiple, interacting variables" (p. 694). However, different numbers have been suggested for acquiring different aspects of vocabulary knowledge, with form recognition needing fewer exposures than grammatical function and meaning recall (e.g., van Zeeland &

Schmitt, 2013a; Webb, 2007). Research into reading (e.g. Webb, 2007) has indicated that 10+ exposures might be needed to accelerate vocabulary learning. Vidal's (2011) findings showed that listening requires more exposure frequencies to words than reading for considerable vocabulary learning to occur. Research on listening has shown that 10 encounters might not be enough for developing and retaining vocabulary knowledge. Brown et al. (2008) suggested that 20+ encounters may be required for vocabulary items to be learned incidentally through listening, a claim supported by van Zeeland and Schmitt (2013a). While they reported fewer exposures (e.g., 7 times) may result in partial development of vocabulary knowledge, over 15 exposures may be required for greater gains and "any meaningful learning of meaning to occur" (p. 621). This has been supported by Pavia et al (2019), where there were greater learning gains for words encountered more frequently (i.e., 18 times) than for words encountered less frequently (i.e., 6 times). Considering the findings of reading studies (e.g., Webb, 2007) and listening studies (e.g., Brown et al., 2008; Pavia et al., 2019), van Zeeland and Schmitt's (2013a) suggestion (i.e., at least 15 times) can be considered a cut-off point where there might be potential for incidental vocabulary learning through listening.

Another line of research into vocabulary learning has been concerned with investigating the potential of a number of discourse types for incidental vocabulary learning, leading to a few corpus-driven studies focusing on low-frequency words in movies (Webb, 2010), television programs (Rodgers & Webb, 2011; Webb & Rodgers, 2009b), and teacher talk (Horst, 2010). However, no corpus-driven research, to the best of our knowledge, has been conducted to explore the potential of both EAL and general-audience podcasts for incidental vocabulary learning.

Webb and Nation (2017) have asserted that knowledge of the most frequent 3,000 word families enables learners to understand different spoken discourse types. Furthermore, Schmitt and Schmitt (2014) have posited that words beyond the most frequent 3,000 word families are necessary for proficient language use and using English for specific purposes. While the most frequent 1,000 word families may receive adequate explicit attention within the limited classroom time (Nation, 2001), the most frequent 2,000 and 3,000 word families may not (Dang & Webb, 2016). Therefore, to achieve the lexical targets of acquiring knowledge of the most frequent 3,000 word families (required to understand most spoken discourse types) and the most frequent 8,000-9,000 word families (necessary to understand most written discourse types), intentional vocabulary learning has to be accompanied by incidental learning (Webb, 2020). Thus, it may be sensible to explore podcasts' potential for the incidental learning of words from the 2,000- and 3,000-word levels as well as beyond to investigate their suitability as a source of incidental vocabulary learning.

Language Learning Through Podcasts

Research on using podcasts in language learning and teaching has attested to their value as an effective tool and their potential for developing language skills (Heilesen, 2010), as they provide authentic aural input (McBride, 2009). Research has also explored the use of podcasts as an extensive listening tool. Chang and Millet (2013, p. 31) defined extensive listening as "doing a lot of easy, comprehensible, and enjoyable listening practice". Alm (2013) examined students' outside-the-classroom listening habits, where

they expressed their attitudes towards their podcast-based learning experiences. The findings showed that the students preferred selecting their own favourite podcasts as this helped them achieve their individual listening objectives at their desired pace of learning. Yeh (2014) investigated university students' podcast use for practicing listening, where they were asked to present their views on podcasts as a language learning device. The results showed that listening to podcasts provided considerable exposure to authentic linguistic input with topics of both personal and academic interests.

Podcasts have been suggested as a useful tool for vocabulary learning (Meier, 2015). Putman and Kingsley (2009) investigated vocabulary learning from listening to podcasts. Half of the participants had access to the podcasts as in-class supplementary material as well as the textbook, whereas the other half did not. The findings indicated that there was a greater improvement in the vocabulary knowledge of the podcast group than that of the non-podcast one. Lu (2007) explored vocabulary learning from podcast listening in a case study of a learner listening to one podcast per week. The learner was asked to transcribe the podcasts while listening, and was then provided with his errors and required to listen to the podcasts again to correct the errors. Analysis of the first and final drafts of the transcripts suggested an improvement in his vocabulary knowledge.

Perhaps the most relevant study in the literature to the present study is the one conducted by Nurmukhamedov and Sharakhimov (2021), who investigated the lexical demands of general-audience podcasts. Using Nation's (2017) BNC/COCA word lists to analyze their 1,137,163-token corpus (170 podcast episodes from nine different popular podcasts), they reported that knowledge of the most frequent 3,000 and 5,000 word families plus proper nouns, marginal words, transparent compounds, and acronyms was required to reach 95% and 98% coverage, respectively. Although they explored the lexical demand of an auditory source of linguistic input, they followed the previous lexical profiling studies and based their analysis on 95% and 98% coverage points, as suggested by lexical coverage studies of written input (e.g., Hu & Nation, 2000; Laufer, 1989; Schmitt et al., 2011). This is surprising as research investigating the effects of lexical coverage on listening comprehension has shown that 90% and 95% are the most appropriate lexical coverage points to examine in relation to auditory sources of linguistic input (Bonk, 2000; van Zeeland & Schmitt, 2013b). Furthermore, their study focused on only exploring the lexical profile of general-audience podcasts with no comparison with podcasts produced for language learning and teaching purposes.

It has been suggested that 95% coverage provides sufficient listening comprehension (Van Zeeland & Schmitt, 2013b). This coverage figure has also been considered the point where there is potential for incidental vocabulary learning from listening (Van Zeeland & Schmitt, 2013a). Van Zeeland and Schmitt (2013b) have asserted that to reach this amount of coverage in listening learners may have to know around 2,000–3,000 word families. This might provide some indication that for podcasts to be considered a useful and appropriate source of auditory linguistic input and vocabulary learning, listeners might need knowledge of between 2,000 and 3,000 word families as this level of lexical knowledge may provide “an appropriate learning goal for unassisted listening” (Webb, 2021, p. 285). However, if podcasts are to be considered authentic auditory materials to enhance vocabulary knowledge, there is a need to explore their lexical load before assigning them to learners. Therefore, this study will attempt to bridge the gap by addressing the following research questions:

1. How many word families are required to understand the vocabulary in podcasts?
2. How is the coverage of EAL podcasts different from that of general-audience ones?
3. To what extent do EAL and general-audience podcasts hold potential for incidental vocabulary learning?

Methodology

Corpus

The transcripts of 543 podcast episodes from eight different programs constituted two corpora comprising an aggregate of 1,188,512 tokens. The total running time was 138 hours, 17 minutes, and 2 seconds (excluding 57 minutes of commercials), representing a relatively large amount of listening time (see Table 2). The podcasts were selected according to three factors: popularity, transcript availability, and a broad range of subjects. The goal was to select podcasts that were well-known, popular, easily accessible, and representative of the taste of a great range of listeners. The popularity criterion was met according to an internet-based podcast service platform called *Chartable* (see <https://chartable.com>). This is run by a podcast analytics and attribution company that provides podcasters and podcast enthusiasts with valuable data about various podcasts available on different platforms. Such data, for example, include charts providing information about top podcasts ranked based on their popularity. As of May 2022, all of the podcasts used in this study were among the top 100 podcast shows based on the data from *Chartable*. Another criterion to consider was the availability of transcripts as the majority of podcasts available on the Internet do not provide any transcripts and only their audio files are accessible. We also ensured that the podcasts were selected from a wide range of topics, including everyday life, science, history, and technology.

The EAL podcasts analyzed in this study contained 445 podcasts (594,292 tokens/74.76 hours) from four programs (*British Council Learn-English Podcasts*, *ESL Pod*, *BBC 6 minute English*, and *VOA Learning English*), covering a range of proficiency levels (i.e., pre-intermediate, intermediate, and upper-intermediate). One criticism levelled at EAL podcasts regards their authenticity as they are modified for pedagogical purposes. However, modifications like shorter sentence length, slower speech rate, and yet use of natural intonation as far as possible make such podcasts more suitable for language learners (Nurmukhamedov & Sadler, 2011). Furthermore, Nurmukhamedov and Sadler (2011) assert that a large number of EAL podcasts (what they referred to as ESL-focused and ESL-super) are prepared by expert users of English; therefore, they are characterized by features such as contextualization of language use and interactions between two or more people, making their content to some extent authentic. On the other hand, the general-audience corpus contained 98 podcasts (594,220 tokens/63.51 hours) from four programs: *In Our Time*, *the Reith Lectures*, *On Being with Krista Tippett*, and *99% Invisible*. No proficiency level is specified for these podcasts, as they are produced without any language learning and teaching purposes. While EAL podcasts are usually used only by language learners and teachers, general audience podcasts may be accessed by those both with and without language learning and teaching purposes. Furthermore, due to podcasts' increasing

popularity, learners with different education levels in different language learning contexts might show interest in using them in their language learning processes.

Table 2

General Information About the Podcasts

Podcasts	Number of Episodes	Number of Tokens	Running Time
In Our Time	37	221,051	20:52:49
The Reith Lectures	19	147,854	15:11:07
On Being with Krista Tippett	16	113,676	13:34:08
99% Invisible	26	111,639	13:52:47
BBC 6 minute English	162	168,272	16:12:00
ESL Pod	93	204,006	29:24:30
VOA Learning English	130	77,025	11:39:56
British Council Learn-English Podcasts	60	144,989	17:29:45
Overall	543	1,188,512	138:17:02

Care was taken to maintain a balance between the two corpora regarding their number of tokens, and the difference was kept to a minimum (only 72 tokens). Despite this small difference, the corpora greatly differed regarding their running times (almost 11 hours). One reason for this might be that words and sentences in EAL podcasts were spoken at a slow pace to make following along easier for learners. Words not spoken in the podcasts (e.g., speakers' names and stage commands) were removed from the transcripts. Following the methodology of previous lexical profiling research (e.g., Nurmukhamedov & Sharakhimov, 2021), we altered connected speech (e.g., *coulda*), contractions (e.g., *they're*), hyphenated words (e.g., *script-writer*), and apostrophized abbreviations (e.g., *makin'*) to conform to the standard spellings found in the BNC/COCA lists. If we had not changed the spellings of such items, the software used to analyze the corpora would have classified them as words *not in the lists*. There were also a number of tokens in the two corpora which the BNC/COCA word lists did not account for, but their family members were found in the lists. For example, the word family with the headword *flourish* and its other members *flourished*, *flourishes*, and *flourishing* can be found in list 4, but not the word *flourishingly* (with a frequency of one in the corpora). We inspected the *not in the lists* results to find such misclassified words. These words were then reclassified and manually added to their respective word families. It should also be noted that despite the large number of entities (more than 22,000) in the BNC/COCA proper nouns list, there can still be a large number of proper nouns in a corpus this list does not account for. Thus, the proper nouns that were in the corpora but not in the BNC/COCA proper nouns list were added to this list. Furthermore, there were a small number of marginal words (e.g., *achoo*, *yikes*) found in the *not in the lists* category which we manually added to list BNC/COCA

list of marginal words (i.e., list 32). It should also be mentioned the transcripts were checked for misspellings. For more information about the podcasts, see Appendix 1.

Analysis

The AntWordProfiler software (Anthony, 2021), version 1.5.1w, was used to analyze the corpora. This program ranks the words occurring in a text according to their frequency and provides results regarding the number and percentage of tokens, types, and word families from each 1,000-word list. Furthermore, this program provides lists of the most frequent word families occurring in each word list. In the present study, the BNC/COCA word lists were used to analyze the lexical demands of the corpora. These comprise 25 word lists of 1,000 word families and four additional lists: proper nouns (e.g., *Damian*), marginal words (e.g., *um*), transparent compounds (e.g., *backdoor*), and acronyms (e.g., *CD*). A word family comprises a headword often accompanied by a number of family members. It is assumed that family members can be understood from knowledge of the stem and affixes (Bauer & Nation, 1993). The word families are rated based on Bauer and Nation's (1993) level-6 classification, which involves inflections and derivational affixes. Words not found in the 29 lists are classified by the software as *not in the lists*. These are items that the BNC/COCA word lists do not account for (e.g., less frequent, specialized, and non-English words), and are normally excluded from the analyses in lexical profiling studies (e.g., Nurmukhamedov & Sharakhimov, 2021).

In keeping with previous studies reporting adequate coverage for comprehending aural input between 90% and 95%, this study also used the same thresholds as the lower and upper boundaries for comprehending podcasts. These thresholds are suggested as the coverage points required to reach adequate comprehension of auditory discourse (van Zeeland & Schmitt, 2013b) in alignment with another study (Bonk, 2000).

Results

Table 3 shows the first 1,000-word level accounted for most of the words, where there was a consistent decrease in the number of tokens as we move towards the bottom of the table (except for the 18,000- and 22,000-word levels where there was an increase in the number of tokens). The first 1,000-word level made up 1,000,546 (84.18%) of the words. However, the most frequent 2,000 word families accounted for 67,042 of the tokens (5.64%), whereas the third 1,000-word level made up only 34,229 of the tokens (2.88%). All lists beyond the third frequency band each accounted for less than 1% of the tokens with the coverage decreasing to less than 0.1% by the 10th level (except for the 18,000-word level with 0.11% coverage), showing the relative importance of knowing the most frequent 3,000 word families. It should also be mentioned that the 4,000-word level still accounted for a large number of word families (961). However, the difference was in the number of times word families in this list were encountered, which was less than the first three 1,000-word levels.

Table 3 also indicates that 3.05% of the tokens were proper nouns, the third highest percentage after the first two frequency bands. Furthermore, marginal words, transparent compounds, and acronyms accounted for 0.71%, 0.39%, and 0.10% of the tokens, respectively. Nation (2006) suggested that marginal words and proper nouns carry a

relatively little lexical burden. Moreover, it is conceivable to assume that the learning burden of transparent compounds might be minimal because they can also be perceived as known through knowledge of their high-frequency parts (Nation, 2016). Furthermore, Nation (2016) argued that learning acronyms and recalling their full form may be easy due to the clues their initial letters provide. Therefore, it is plausible to add their coverage points to the coverage required to know each 1,000-word list. It should also be mentioned that 0.07% of the words in the corpora were not found in any of the BNC/COCA word lists. These were less frequent, specialized, or non-English words (e.g., *peroxisome*, and *siniora*). These were excluded from the analysis, and therefore their coverage was not taken into account.

Table 3*Tokens and Word Families at Each Word Level for all Podcasts*

Word List	Tokens		Word Family
	Raw	%	
1,000	1000546	84.18	1000
2,000	67042	5.64	992
3,000	34229	2.88	998
4,000	11730	0.99	961
5,000	7165	0.60	888
6,000	4413	0.37	763
7,000	2989	0.25	667
8,000	2109	0.18	583
9,000	1606	0.14	458
10,000	960	0.08	368
11,000	816	0.07	289
12,000	599	0.05	213
13,000	428	0.04	171
14,000	337	0.03	123
15,000	245	0.02	108
16,000	228	0.02	83
17,000	109	0.01	53
18,000	1294	0.11	50
19,000	69	0.01	35
20,000	51	0.00	29
21,000	32	0.00	20
22,000	63	0.01	19
23,000	31	0.00	17
24,000	16	0.00	9
25,000	31	0.00	14
proper nouns	36241	3.05	4504
marginal words	8493	0.71	38
transparent compounds	4646	0.39	791
acronyms	1145	0.10	99
not in the lists	849	0.07	443
total	1188512		14786

Table 4 illustrates the cumulative coverage for the tokens with proper nouns (PN), marginal words (MW), transparent compounds (TC), and acronyms (AC). Regarding the first research question, the results showed that knowledge of the most frequent 1,000 word families plus PN/MW/TC/AC did not provide 90% coverage for all podcasts. To reach this coverage point, however, knowledge of at least 2,000 word families was required. Furthermore, with knowledge of 3,000 word families plus PN/MW/TC/AC, 96.95% of the words would be known.

Table 4

Cumulative Coverage (Including Proper Nouns, Marginal Words, Transparent Compounds, and Acronyms) for all Podcasts

word list	cumulative coverage
1,000	88.43
2,000	94.07 ^α
3,000	96.95 ^β
4,000	97.94
5,000	98.54
6,000	98.91
7,000	99.16
8,000	99.34
9,000	99.48
10,000	99.56
11,000	99.63
12,000	99.69
13,000	99.73
14,000	99.76
15,000	99.78
16,000	99.80
17,000	99.81
18,000	99.92
19,000	99.93
20,000	99.93
21,000	99.93
22,000	99.94
23,000	99.94
24,000	99.94
25,000	99.94

Note. α: reaching 90%; β: reaching 95%

Considering the second research question, Table 5 indicates that vocabulary sizes of 1,000 and 2,000 word families plus PN/MW/TC/AC were required to reach 90% coverage in the EAL and general-audience podcasts, respectively. Moreover, reaching 95% coverage also needed different vocabulary sizes in the corpora, where knowledge of 2,000 (EAL) and 3,000 (general-audience) word families plus PN/MW/TC/AC was needed. Overall, there is evidence that EAL podcasts may be lexically less demanding than

general-audience ones as the results show they may need smaller vocabulary sizes to reach 90% and 95% coverage.

Table 5

Cumulative Coverage (Including Proper Nouns, Marginal Words, Transparent Compounds, and Acronyms) for Each Corpus

word list	EAL	General-Audience
1,000	90.14 ^α	86.73
2,000	95.42 ^β	92.73 ^α
3,000	97.32	96.59 ^β
4,000	98.14	97.74
5,000	98.65	98.44
6,000	98.99	98.84
7,000	99.18	99.16
8,000	99.33	99.37
9,000	99.47	99.50
10,000	99.53	99.60
11,000	99.58	99.69
12,000	99.63	99.74
13,000	99.66	99.78
14,000	99.69	99.81
15,000	99.70	99.84
16,000	99.71	99.87
17,000	99.72	99.88
18,000	99.92	99.90
19,000	99.92	99.91
20,000	99.92	99.91
21,000	99.92	99.91
22,000	99.93	99.92
23,000	99.93	99.92
24,000	99.93	99.92
25,000	99.93	99.92

Note. α: reaching 90%; β: reaching 95%

It is also important to note that the corpora contained a large number of episodes (i.e., 543), and each episode may need different vocabulary sizes to reach the required coverage levels. To address this, we conducted a separate analysis of the eight different programs used in this study and also one random episode from each. Table 6 shows that there was some variation in the vocabulary sizes needed to reach 90% and 95% coverage among different programs of the same category of podcasts. For example, our earlier findings revealed that to achieve 90% and 95% coverage in EAL podcasts vocabulary sizes of 1,000 and 2,000 word families plus PN/MW/TC/AC were required, respectively. However, our analysis showed that while two of the programs within this category required the same vocabulary targets as the whole category to reach the same coverage levels, the other two did not. Accordingly, while the podcasts from *British Council* and *ESL Pod* required 1,000 and 2,000 word families plus PN/MW/TC/AC to reach 90% and 95%

coverage, the other two programs within the category of EAL podcasts (i.e., *BBC 6 Minute English* and *VOA Learning English*) needed 2,000 and 3,000 word families plus PN/MW/TC/AC to achieve the same coverage points. Variation in the vocabulary targets was also indicated upon analysis of one random episode from each program. For example, the results demonstrated that the vocabulary sizes necessary to reach 90% and 95% coverage in the episode *Allergic to Cats* from *ESL Pod* were 2,000 and 3,000 word families including PN/MW/TC/AC, respectively. These were different vocabulary sizes from the ones required for the whole program and the EAL category of podcasts analyzed in this study.

Table 6

Cumulative Coverage (Including Proper Nouns, Marginal Words, Transparent Compounds, and Acronyms) for Different Programs and One Random Episode From Each

General-audience	99% Invisible	one	84.81	90.78 ^α	94.06	94.88	97.93 ^β	98.36	98.63	98.90	99.17	99.33	99.41	99.43	99.51	99.51	99.51	99.51	99.55	99.59	99.71	99.71	2561	
		all	86.96	92.76 ^α	95.97 ^β	97.14	98.16	98.65	98.95	99.16	99.31	99.45	99.57	99.66	99.71	99.73	99.76	99.77	99.79	99.81	99.81	99.82	111,639	
	On Being with Krista Tippett	one	90.71 ^α	94.82	97.58 ^β	98.24	98.61	98.67	99.08	99.22	99.36	99.47	99.80	99.82	99.82	99.83	99.83	99.85	99.85	99.87	99.87	99.89	99.89	8379
		all	88.62	93.85 ^α	96.95 ^β	98.04	98.65	99.03	99.32	99.47	99.59	99.63	99.72	99.76	99.78	99.79	99.80	99.82	99.82	99.85	99.86	99.86	113,676	
	The Reith Lectures	one	79.16	88.42	96.62 ^β	97.72	98.53	98.81	99.23	99.44	99.63	99.79	99.86	99.90	99.94	99.94	99.94	99.94	99.98	99.98	99.98	99.98	99.98	5705
		all	85.86	92.50 ^α	96.83 ^β	97.96	98.58	98.92	99.22	99.42	99.57	99.67	99.75	99.80	99.84	99.87	99.89	99.91	99.92	99.93	99.93	99.93	147,854	
	In Our Time	one	84.05	88.80	94.86 ^α	95.80 ^β	96.99	97.59	98.23	98.39	98.80	98.90	98.93	99.00	99.25	99.36	99.41	99.60	99.74	99.78	99.81	99.81	7305	
		all	86.20	92.26 ^α	96.52 ^β	97.71	98.34	98.75	99.10	99.34	99.45	99.56	99.64	99.68	99.73	99.77	99.81	99.86	99.87	99.88	99.89	99.89	221,051	
	EAL	ESL Pod	one	89.50	92.90 ^α	94.86	95.87 ^β	96.98	97.19	98.25	98.67	98.67	98.78	98.78	98.83	98.83	98.83	98.83	98.83	99.36	99.36	99.36	1885	
		ESL Pod VOA	all	90.98 ^α	95.58 ^β	97.14	97.92	98.53	98.85	99.07	99.24	99.39	99.44	99.51	99.56	99.60	99.63	99.64	99.65	99.65	99.95	99.95	99.95	204,006
		Learning English VOA	one	93.63 ^α	96.82 ^β	98.62	98.90	98.90	98.90	99.18	99.46	99.46	99.46	100	-	-	-	-	-	-	-	-	-	722
		Learning English BBC 6-minute English	all	86.27	94.40 ^α	97.68 ^β	98.57	98.93	99.19	99.39	99.53	99.63	99.71	99.74	99.82	99.86	99.88	99.89	99.90	99.91	99.93	99.93	99.93	77,025
BBC 6-minute English		one	92.92 ^α	95.19 ^β	97.46	98.00	98.00	98.27	98.27	98.36	98.45	98.72	98.72	98.72	98.72	98.72	98.72	98.72	98.72	98.81	98.81	98.81	1102	
BBC 6-minute English British Council podcasts		all	87.93	94.23 ^α	96.97 ^β	97.94	98.47	98.95	99.20	99.38	99.56	99.64	99.70	99.74	99.76	99.80	99.82	99.83	99.85	99.86	99.87	99.88	168,272	
British Council podcasts		one	95.58 ^β	98.27	98.92	99.18	99.40	99.56	99.56	99.58	99.62	99.62	99.66	99.66	99.66	99.66	99.66	99.66	99.66	99.98	99.98	99.98	4938	
word level		all	93.62 ^α	97.16 ^β	97.84	98.50	98.90	99.15	99.21	99.28	99.38	99.42	99.44	99.49	99.50	99.51	99.53	99.53	99.53	99.92	99.92	99.92	144989	
		tokens	1,000	2,000	3,000	4,000	5,000	6,000	7,000	8,000	9,000	10,000	11,000	12,000	13,000	14,000	15,000	16,000	17,000	18,000	19,000	20,000	tokens	

Note. α: reaching 90%; β: reaching 95%

The final research question addressed the potential of each type of podcast for incidental vocabulary learning. As the results showed, knowledge of 1,000–3,000 word families was required to reach 90%–95% coverage in the two corpora. This may suggest that the most likely words to be learned incidentally by learners might be those in the 2,000- and 3,000-word levels. Furthermore, we also investigated the potential for the incidental learning of words in levels beyond the 3,000-word level (i.e., 4,000–25,000-word levels). Table 7 shows the frequency of occurrence of word families in 2,000-, 3,000- and 4,000–25,000-word levels. Accordingly, almost 60% of the words from the 2,000-word level had a frequency of 15+ (e.g., *culture, tradition, value, evidence*) in both EAL and general audience podcasts. Furthermore, results showed that nearly 29% (e.g., *colleague, constitution*) and 50% (e.g., *fundamental, capacity*) of the word families from the third 1,000-word level were encountered 15+ times in the EAL and general-audience podcasts, respectively. The results also indicated that only 176 (e.g., *census, spendthrift*) of the 2304 word families encountered in the 4,000–25,000-word levels in the EAL podcasts had an exposure frequency of 15+, while the figure for general-audience podcasts was 214 (e.g., *commodity, solitude*) (out of 5,152 word families). For an example of a transcript highlighting words from frequency levels, see Appendix 2.

Table 7

Number and Percentage of Occurrence of Word Families in the 2,000-, 3,000, and 4,000–25,000-word levels in Each Corpus

occurrence	EAL						general audience					
	2,000		3,000		4,000–25,000		2,000		3,000		4,000–25,000	
	WF	%	WF	%	WF	%	WF	%	WF	%	WF	%
1	29	3.01	97	11.26	689	29.90	27	2.78	17	1.71	2088	40.52
2	31	3.21	77	8.94	314	13.62	26	2.68	35	3.53	974	18.90
3-4	53	5.50	132	15.33	443	19.22	61	6.30	84	8.49	921	17.87
5-7	101	10.48	145	16.84	394	17.10	102	10.53	125	12.63	584	11.33
8-10	76	7.89	84	9.75	157	6.81	87	8.98	119	12.03	218	4.23
11-14	100	10.38	78	9.05	131	5.68	80	8.26	118	11.93	153	2.96
15+	573	59.50	248	28.80	176	7.63	585	60.43	491	49.64	214	4.15
total WFs	963		861		2304		968		989		5152	

Note. WF: word family

Discussion and Pedagogical Implications

How Many Word Families Are Required to Understand the Vocabulary in Podcasts?

As for the first research question, our findings showed that vocabulary knowledge of 2,000 word families plus PN/MW/TC/AC was required to reach 90% coverage in podcasts. Our results also confirmed those of Nurmukhamedov and Sharakhimov (2021), suggesting that 3,000 word families plus PN/MW/TC/AC may be required for 95% coverage. The results also indicated that the four additional lists accounted for a relatively significant amount of coverage (4.25%), highlighting their importance for gaining adequate comprehension. This is particularly true about proper nouns as they provided the third highest amount of coverage, showing their relative significance for comprehension while listening to podcasts.

Using podcasts in language teaching carries important pedagogical implications. While 98% coverage may be the ideal coverage for very high listening comprehension, 95% is still sufficient for reasonable comprehension to occur (van Zeeland & Schmitt, 2013b). For teachers, the results are instructive because the lexical load is an abiding concern as they seek engaging materials. Teachers should pay attention to the vocabulary demand of a podcast because understanding all the words might be difficult for learners when listening (Meier, 2015), leading to an unpleasant learning experience through listening to an unsuitable podcast (Alm, 2013). However, as Webb and Rodgers (2009a) rightly point out with respect to movies, aural materials like podcasts can serve as an effective supplement to written texts which may form the bulk of their instructional materials. Receptive understanding of the first three 1,000-word levels could be initially aimed at, allowing podcasts to become instructional materials at 95% coverage. Moreover, learners should know 3,000 word families if they want to use auditory linguistic materials like podcasts to further boost their linguistic knowledge.

How Is the Coverage of EAL Podcasts Different From That of General-Audience Ones?

Regarding the second research question, our findings showed that EAL podcasts were lexically less demanding than general-audience ones. This is not surprising because the latter are not created for language learning and teaching purposes, and there is, therefore, no control exercised over their lexical density and breadth, leading to the occurrence of more lower-frequency vocabulary. Our findings regarding lexical demands and rate of speech clearly showed that the aim of creating EAL podcasts is for them to be easily understood, and clearly care is taken to limit their lexical burden. We should also bear in mind that the vocabulary learning targets determined in the present study to achieve 90% and 95% coverage of each category of podcasts were the mean number of word families and that there might be a great deal of variation among different episodes within the same program and also the same category of podcasts. This has been rightly addressed by Webb (2021) asserting that “the learning target that is indicated by the analysis of a corpus will not always reflect the vocabulary knowledge needed to understand individual texts for that discourse type” (p. 286).

It is also useful to draw a comparison between both types of podcasts and other auditory discourse types regarding lexical load to obtain a sense of podcasts' suitability to be integrated into a listening program. According to our findings, the vocabulary load of EAL podcasts is lower than that of most of the auditory discourse types. However, two discourse types, namely teacher-selected songs (Tegge, 2017) and teacher talk (Horst, 2010), have been suggested to be similar to EAL podcasts regarding lexical demand. These have been suggested to be lexically less demanding than other auditory discourse types. On the other hand, general-audience podcasts are similar to a great extent to other auditory genres regarding both 90% and 95% coverage levels, requiring 2,000 and 3,000 word families, respectively. However, when the aim is 95% coverage, there are three exceptions that have been suggested to be more lexically demanding than general-audience podcasts: rap songs (Tegge & Coxhead, 2021), academic lectures (Dang & Webb, 2014), and CanTEST listening (Webb & Paribakht, 2015).

This difference between the two types of podcasts could be exploited by teachers. They could start with EAL ones, giving learners a head start with more frequent word families, which is supported by Nurmukhamedov and Sharakhimov (2021), suggesting that listening to a podcast with a lower lexical demand might ease the comprehension burden. Gradually as they progress in proficiency and vocabulary knowledge, they can be exposed to general-audience podcasts, boosting their vocabulary size. Furthermore, teachers and learners can benefit from the comparison between podcasts and different spoken discourse types as such a comparison reveals that podcasts, both EAL and general-audience, are somewhere in the middle of the lexical demand continuum, and can therefore be considered a valuable source of mid-level authentic materials to improve listening comprehension and vocabulary learning.

To What Extent Do EAL and General-Audience Podcasts Hold Potential for Incidental Vocabulary Learning?

Considering the final research question, the results demonstrated that listening to both types of podcasts might hold relatively great potential for the incidental learning of words from the 2,000-word level. This is the case as the majority of word families (almost 60%) from this word level were encountered frequently enough (i.e., 15+ times) in both corpora. Furthermore, our findings showed that general-audience podcasts may have greater potential for the incidental learning of words from the third 1,000-word level compared with EAL podcasts. This is because the number of word families encountered 15+ times in the general-audience podcasts was almost twice that in the EAL podcasts. Our results also indicated that listening to podcasts, both EAL and general-audience, may not hold great potential for the incidental learning of words beyond the 3,000-word level as a very small number of such words (nearly 8% in the EAL and 5% in the general-audience podcasts) were encountered frequently enough. This means that a large number of such words are unlikely to be learned incidentally within almost 139 hours of listening time. However, if listening time were increased, then podcasts' potential for the incidental learning of such words might also increase because individuals might spend more time listening to podcasts than the amount of listening time analyzed in this study.

Though podcasts' potential for incidental vocabulary learning was addressed concerning exposure frequency in this study, there are other factors that might influence

this process, including L2 proficiency (Zahar et al., 2001), background knowledge (Pulido, 2004), and learners' prior lexical knowledge (Webb & Paribakht, 2015). For example, if learners lack adequate knowledge of high-frequency words, they might not give their attention to lower-frequency ones however frequent they are. This suggests that even if a word is encountered a certain number of times, it does not necessarily mean it can be learned incidentally. Therefore, caution should be exercised when interpreting the results.

As for the pedagogical implications, listening to podcasts is suggested to hold great potential for incidental vocabulary learning (Meier, 2015), providing support for creating extensive listening programs through podcasts. This provides a great opportunity for repeated exposure to unknown vocabulary, paving the path for incidental vocabulary learning. High exposure through listening may be a necessary condition for words to be learned, and podcasts can be a vehicle for fulfilling this condition. One podcast is hardly enough to ensure much incidental vocabulary learning as the unknown words are unlikely to be heard enough. To enhance the chances of incidental vocabulary learning through podcasts, the most promising solution might be increasing the amount of listening time, which is perfectly possible in most language learning settings given the easy access to the Internet. One strategy would be listening to several podcasts so unknown words are repeated enough to make incidental learning possible. Another strategy is listening to an episode multiple times, as also suggested by Pavia et al. (2019) for songs, because this provides an opportunity for listeners to be exposed to words several times, increasing the chance of incidental learning. Despite the necessity of increased listening time for fully learning words, a few encounters or even single encounters, as shown for extensive reading (Freimuth, 2020), are not completely futile, laying the foundation for a partial grasp of word knowledge (Webb, 2007).

Regarding which vocabulary items should be prioritized, Vilkaitė-Lozdienė and Schmitt (2020) asserted that the most useful words are the high-frequency vocabulary, and therefore require special attention. Our findings confirmed those of Dang and Webb (2016) and Dang et al. (2020) regarding high-frequency vocabulary as we found that the most frequent 1,000–2,000 word families provided 90% coverage of podcasts. Therefore, it seems plausible to assume that such words should be given particular attention in an extensive listening program through podcasts. One approach might be increasing exposure to podcasts on a wide range of topics.

Despite the higher likelihood of encountering high-frequency word families, we cannot plausibly dismiss the importance of lower-frequency vocabulary, especially to those learning English for specific purposes (Vilkaitė-Lozdienė & Schmitt, 2020). Following previous research (e.g., van Zeeland, 2018; Webb & Rodgers, 2009a) suggesting pre-teaching unknown, key, and lower-frequency words to reduce comprehension burden, we also recommend employing this strategy for lower-frequency words in preparing students for listening to podcasts as most of such words are unlikely to be learned incidentally. Moreover, the focus of many materials is topic-based, meaning lower-frequency words regularly receive explicit attention because they are considered essential for understanding certain discourse (Vilkaitė-Lozdienė & Schmitt, 2020).

A further extension of the pedagogical potential of podcasts regards known words. While knowing the forms and meanings of words is undoubtedly crucial in developing competence in an L2, it is equally important to embed this knowledge within authentic webs of links to other words to facilitate fluent performance (Wood, 2010). In this respect,

podcasts present ample opportunity to consolidate already known words, affording the development of collocational links and other vocabulary associations (Wray, 2008), which is particularly true for advanced learners past the milestone of 5,000 word families.

Although the rationale behind establishing extensive listening programs through podcasts is evident, from a pedagogical viewpoint, listening to podcasts for language learning is still an under-explored avenue. Therefore, using podcasts for pedagogical purposes requires attention to some considerations. Learners' listening habits should be considered because understanding such habits can yield better pedagogical practices. Little is known about how often learners listen to podcasts outside the classroom. Nor is much known about which programs they choose to listen to. Furthermore, the relationship between out-of-class listening and classroom activities should be considered. The incidental vocabulary potential highlighted in this study may be enhanced through such activities as group discussion based on episodes (listened to outside of the classroom) and focus-on-form activities concentrating on specific segments of each podcast.

Limitations

This study has a number of limitations that should be considered when interpreting the findings. First, we should not forget vocabulary is one of manifold factors affecting the comprehension of podcasts. Other factors such as syntactic complexity and idiomatic use of vocabulary are among potentially significant mediators of comprehension. Therefore, knowing all the words in a passage does not necessarily lead to understanding the intended message (Graham, 2006). Although the vocabulary learning goals determined by lexical profiling studies are useful, achieving those goals does not guarantee that the discourse type in question will necessarily be understood, showing that the vocabulary targets associated with such studies might not be meaningful and valid (Webb, 2021). This means that the vocabulary learning targets indicated in the present study do not necessarily ensure comprehension of podcasts. Furthermore, although there are arguments both for (e.g., Laufer, 2021; Snoder & Laufer, 2022) and against (e.g., Brown et al., 2021; McLean, 2017) using word families as the unit of analysis in vocabulary research, we utilized word families in our analysis for comparability reasons to the previous lexical profiling studies that adopted the same approach. We also noted that frequency of occurrence is one of the many factors influencing incidental vocabulary learning. Furthermore, the division of podcasts into EAL and general-audience ones and the concrete differences that came to light in the analysis should not blind us to the uniqueness of each podcast and its lexical load. The podcasts analyzed in this study came from a variety of subgenres, each representing a different semantic domain with implications for the vocabulary covered.

Future research could address some of these limitations. For example, research can explore the degree to which learners who have reached the vocabulary learning goals determined in our study can understand podcasts. This will help clarify the extent to which the vocabulary learning targets we have indicated are meaningful and valid. Furthermore, following the methodology used by van Zeeland and Schmitt (2013a, 2013b), a potential direction for future research can be investigating the relationship between learners' vocabulary knowledge, their understanding of podcasts and incidental vocabulary learning.

Correspondence should be addressed to Masoud Motamedynia.
Email: m.motamedynia@alumni.ut.ac.ir

Declaration of Interest Statement

No potential competing interest was reported by the authors.

Data Availability Statement

The data that support the findings of this study are available from the corresponding authors, upon request.

References

- Alm, A. (2013). Extensive listening 2.0 with foreign language podcasts. *Innovation in Language Learning and Teaching*, 7(3), 266–280.
<https://doi.org/10.1080/17501229.2013.836207>
- Anthony, L. (2021). AntWordProfiler (Version 1.5.1) [Computer Software]. Tokyo, Japan: Waseda University. <https://www.laurenceanthony.net/software>
- Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253–279. <https://doi.org/10.1093/ijl/6.4.253>
- Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening*, 14(1), 14–31.
<https://doi.org/10.1080/10904018.2000.10499033>
- Brown, D., Stewart, J., Stoeckel, T., & Mclean, S. (2021). The coming paradigm shift in the use of lexical units. *Studies in Second Language Acquisition*, 43, 950–953.
<https://doi.org/10.1017/S0272263121000668>
- Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, 20(2), 136–163. <https://files.eric.ed.gov/fulltext/EJ815119.pdf>
- Cervetti, G., & Wright, T. (2020). The Role of Knowledge in Understanding and Learning From Text. In E. Birr Moje, P. Afflerbach, P. Enciso, & N. Lesaux (Eds.), *Handbook of Reading Research* (pp. 237–260). Routledge.
- Chang, A., & Millett, S. (2013). The effect of extensive listening on developing L2 listening fluency: some hard evidence. *ELT Journal*, 68(1), 31–40.
<https://doi.org/10.1093/elt/cct052>
- Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics*, 31(5), 693–713.
<http://doi.org/10.1093/applin/amq031>
- Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, 33(1), 66–76.
<https://doi.org/10.1016/j.esp.2013.08.001>
- Dang, T. N. Y., & Webb, S. (2016). Evaluating lists of high-frequency words. *ITL-International Journal of Applied Linguistics*, 167(2), 132–158.
<https://doi.org/10.1075/itl.167.2.02dan>
- Dang, T. N. Y., Webb, S., & Coxhead, A. (2020). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, 26(4), 617–641. <https://doi.org/10.1177/1362168820911189>

- Douglas, S. R. (2013). The lexical breadth of undergraduate novice level writing competency. *Canadian Journal of Applied Linguistics*, 16(1), 152–170. <http://journals.hil.unb.ca/index.php/CJAL/article/view/21176>
- Douglas, S. R. (2016). The relationship between lexical frequency profiling measures and rater judgements of spoken and written general English language proficiency on the CELPIP-General Test. *TESL Canada Journal*, 32, 43–64. <https://doi.org/10.18806/tesl.v32i0.1217>
- Durbahn, M., Rodgers, M., & Peters, E. (2020). The relationship between vocabulary and viewing comprehension. *System*, 88, 102–166. <https://doi.org/10.1016/j.system.2019.102166>
- Feng, Y., & Webb, S. (2019). Learning vocabulary through reading, listening, and viewing. *Studies in Second Language Acquisition*, 42(3), 499–523. <https://doi.org/10.1017/s0272263119000494>
- Freimuth, H. (2020). Are academic English words learned incidentally? A Canadian case study. *BC TEAL Journal*, 5(1), 32–43. <https://doi.org/10.14288/bctj.v5i1.344>
- Graham, S. (2006). Listening comprehension: The learners' perspective. *System*, 34(2), 165–182. <https://doi.org/10.1016/j.system.2005.11.001>
- Heilesen, S. (2010). What is the academic efficacy of podcasting? *Computers & Education*, 55(3), 1063–1068. <https://doi.org/10.1016/j.compedu.2010.05.002>
- Horst, M. (2010). How well does teacher talk support incidental vocabulary acquisition? *Reading in a Foreign Language*, 22(1), 161–180. <https://doi.org/10.10125/66646>
- Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–430. https://scholarspace.manoa.hawaii.edu/bitstream/10125/66973/13_1_10125_66973_rfl131hsuehchao.pdf
- Jin, Z., & Webb, S. (2020). Incidental vocabulary learning through listening to teacher talk. *The Modern Language Journal*, 104(3), 550–566. <https://doi.org/10.1111/modl.12661>
- Kaneko, M. (2015). Vocabulary size required for the TOEFL iBT Listening Section. *The Language Teacher*, 39(1), 9–14. <https://doi.org/10.37546/JALTTLT39.1-2>
- Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters
- Laufer, B. (2021). LEMMAS, FLEMMAS, WORD FAMILIES, AND COMMON SENSE. *Studies in Second Language Acquisition*, 43(5), 965–968. <https://doi.org/10.1017/S0272263121000656>
- Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, 22(1), 1–26. <https://doi.org/10.1093/applin/22.1.1>
- Laufer, B., & Sim, D. (1985). An attempt to measure the threshold of competence for reading comprehension. *Foreign Language Annals*, 18(5), 405–411.
- Lu, J. (2007). Podcasting: A fresh solution for old problems. Wireless Ready Symposium E-Proceedings. [Online]. pp. 83–95. <http://opinion.nucba.ac.jp/~thomas/AngLu.pdf>
- Macmillan. (n.d.). Podcast. In *Macmillan Dictionary*. Retrieved February 17, 2021, from <https://www.macmillandictionary.com/dictionary/british/podcast>

- Matthews, J. (2018). Vocabulary for listening: Emerging evidence for high and mid-frequency vocabulary knowledge. *System*, 72, 23–36. <https://doi.org/10.1016/j.system.2017.10.005>
- McBride, K. (2009). Podcasts and second language learning: Promoting listening comprehension and intercultural competence. In L. Abraham & L. Williams (Eds.), *Electronic Discourse in Language Learning and Language Teaching* (pp. 153–168). John Benjamins. <https://doi.org/10.1075/llt.25.12mcb>
- McLean, S. (2017). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, 39(6), 823–845. <https://doi.org/10.1093/applin/amw050>
- Meier, A. (2015). L2 incidental vocabulary acquisition through extensive listening to podcasts. *Studies in Applied Linguistics & TESOL*, 15, 72–84. <https://doi.org/10.7916/D8KP9D2G>
- Nation, I. S. P. (2001). How many high frequency words are there in English? *Language, Learning and Literature: Studies Presented to Hakan Ringbom*. Abo Akademi University, Abo: English Department Publications, 4, 167–181.
- Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer, *Vocabulary in a Second Language* (pp. 3–13). John Benjamins. <https://doi.org/10.1075/llt.10.03nat>
- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1), 59–82. <https://doi.org/10.3138/cmlr.63.1.59>
- Nation, I. S. P. (2012). The BNC/COCA word family lists. <http://www.victoria.ac.nz/lals/about/staff/paul-nation>
- Nation, I. S. P. (2016). *Making and Using Word Lists for Language Learning and Testing*. John Benjamins.
- Nation, I. S. P. (2017). The BNC/COCA level 6 word family lists (Version 1.0.0) [Data file]. <http://www.victoria.ac.nz/lals/staff/paul-nation.aspx>
- Nguyen, C. D. (2020). Lexical features of reading passages in English-language textbooks for Vietnamese high-school students: Do they foster both content and vocabulary gain? *RELC Journal*, 52(3), 509–522. <https://doi.org/10.1177/0033688219895045>
- Nurmukhamedov, U. (2017). Lexical coverage of TED Talks: Implications for vocabulary instruction. *TESOL Journal*, 8(4), 768–790. <https://doi.org/10.1002/tesj.323>
- Nurmukhamedov, U. & Sadler, R. (2011). Podcasts in four categories: Applications to language learning. In M. Abdous & B. Bacer (Eds.), *Academic Podcasting and Mobile Assisted Language Learning: Applications and Outcomes* (pp. 176–195). IGI Global. <https://10.4018/978-1-60960-141-6.ch011>
- Nurmukhamedov, U., & Sharakhimov, S. (2021). Corpus-based vocabulary analysis of English podcasts. *RELC Journal*, Online Advance Publication. <https://doi.org/10.1177/0033688220979315>
- Pavia, N., Webb, S., & Faez, F. (2019). Incidental vocabulary learning through listening to songs. *Studies in Second Language Acquisition*, 41(4), 745–768. <https://doi.org/10.1017/s0272263119000020>
- Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language*, 22(1), 31–55.

- Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through watching a single episode of L2 television. *Studies in Second Language Acquisition*, 40(3), 1–27.
- Phung, D. H., & Ha, H. T. (2022). Vocabulary demands of the IELTS listening test: An in-depth analysis. *SAGE Open*, 12, 1–13. <https://doi.org/10.1177/21582440221079934>
- Pulido, D. (2004). The relationship between text comprehension and second language incidental vocabulary acquisition: A matter of topic familiarity? *Language Learning*, 54(3), 469–523. <https://doi.org/10.1111/j.0023-8333.2004.00263.x>
- Putman, S., & Kingsley, T. (2009). The atoms family: Using podcasts to enhance the development of science vocabulary. *The Reading Teacher*, 63(2), 100–108. <https://doi.org/10.1598/rt.63.2.1>
- Qian, D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52(3), 513–36. <https://doi.org/10.1111/1467-9922.00193>
- Rodgers, M. P. H. & Heidt, J. (2020). Levelling up comprehensible input and vocabulary learning: The lexical profile of video games. In Werner, V. & Tegge, F. (Eds.), *Pop culture in language education*. Routledge.
- Rodgers, M. P. H., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly*, 45(4), 689–717. <https://doi.org/10.5054/tq.2011.268062>
- Rodgers, M. P. H., & Webb, S. (2016). Listening to lectures. In K. Hyland & P. Shaw (Eds.), *The Routledge Handbook of English for Academic Purposes* (pp. 165–176). Routledge.
- Rodgers, M. P. H., & Webb, S. (2020). Incidental vocabulary learning through viewing television. *International Journal of Applied Linguistics*, 171(2), 191–220. <https://doi.org/10.1075/itl.18034.rod>
- Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 47(4), 484–503. <https://doi.org/10.1017/s0261444812000018>
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1), 26–43. <https://doi.org/10.1111/j.1540-4781.2011.01146.x>
- Snoder, P., & Laufer, B. (2022). EFL Learners' Receptive Knowledge of Derived Words: The Case of Swedish Adolescents. *TESOL Quarterly*, Advance Online Publication. <https://doi.org/10.1002/tesq.3101>
- Tegge, F. (2017). The lexical coverage of popular songs in English language teaching. *System*, 67, 87–98. <https://doi.org/10.1016/j.system.2017.04.016>
- Tegge, F. & Coxhead, A. (2021). Exploring the vocabulary of rap lyrics. In V. Werner & F. Tegge (Eds.), *Pop culture in language education: Theory, research, practice* (pp. 71–84). Routledge.
- van Zeeland, H. (2018). Vocabulary in listening. In W. Renandya & G. Hu (Eds.), *The TESOL Encyclopedia of English Language Teaching* (pp. 1–6). Wiley.
- van Zeeland, H., & Schmitt, N. (2013a). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, 41(3), 609–624. <https://doi.org/10.1016/j.system.2013.07.012>

- van Zeeland, H., & Schmitt, N. (2013b). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, 34(4), 457–479. <https://doi.org/10.1093/applin/ams074>
- Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, 61(1), 219–258. <https://doi.org/10.1111/j.1467-9922.2010.00593.x>
- Vilkaitė-Lozdienė, L. & Schmitt, N. (2020). Frequency as a guide for vocabulary usefulness: High-, mid-, and low-frequency words. In S. Webb (Eds.), *The Routledge Handbook of Vocabulary Studies* (pp. 81–95). Routledge.
- Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28(1), 46–65. <https://doi.org/10.1093/applin/aml048>
- Webb, S. (2010). A corpus driven study of the potential for vocabulary learning through watching movies. *International Journal of Corpus Linguistics*, 15(4), 497–519. <https://doi.org/10.1075/ijcl.15.4.03web>
- Webb, S. (2020). Incidental vocabulary learning. In S. Webb (Eds.) *The Routledge Handbook of Vocabulary Studies* (pp.225–239). Routledge.
- Webb, S. (2021). Research investigating lexical coverage and lexical profiling: What we know, what we don't know, and what needs to be examined. *Reading in a Foreign Language*, 33(2), 287–302.
- Webb, S., & Macalister, J. (2013). Is text written for children useful for L2 extensive reading? *TESOL Quarterly*, 47(2), 300–322. <https://doi.org/10.1002/tesq.70>
- Webb, S., & Nation, I.S.P. (2017). *How Vocabulary is Learned*. Oxford: Oxford University Press.
- Webb, S., & Paribakht, T. (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. <https://doi.org/10.1016/j.esp.2014.11.001>
- Webb, S., & Rodgers, M. P. H. (2009a). The Lexical coverage of movies. *Applied Linguistics*, 30(3), 407–427. <https://doi.org/10.1093/applin/amp010>
- Webb, S., & Rodgers, M. P. H. (2009b). Vocabulary demands of television programs. *Language Learning*, 59(2), 335–366. <https://doi.org/10.1111/j.1467-9922.2009.00509.x>
- Winn, R. (2021). 2021 Podcast Stats & Facts (New Research from Apr 2021). <https://www.podcastinsights.com/podcast-statistics/>
- Wood, D. (2010). *Formulaic Language and Second Language Speech Fluency: Background, Evidence and Classroom Applications*. Continuum.
- Wray, A. (2008). *Formulaic Language: Pushing the Boundaries*. Oxford University Press.
- Yeh, C. (2014). An investigation of a podcast learning project for extensive listening. *Language Education in Asia*, 4(2), 135–149. <https://doi.org/10.5746/leia/13/v4/i2/a04/yeh>
- Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *The Canadian Modern Language Review*, 57(4), 541–572. <https://doi.org/10.3138/cmlr.57.4.541>

Appendix 1

Podcasts*		
Features	EAL	General-Audience
Number of Episodes	445	98
Number of Tokens	594,273	594,211
Running Time	74:46:11	63:30:51
Subject Area	technology, science, history, everyday life, nature, world issues, language learning, psychology	science, history, technology, everyday life, psychology, sociology, academia, world issues

*For more information about the podcasts, see <https://chartable.com>

Appendix 2

An Excerpt From an Episode Entitled “*In Praise of Politics*” (*The Reith Lectures*) Showing the Distribution of Words From the 1,000-, 2,000-, **3,000-** and **4,000–25,000** Word Lists, and Proper Nouns (*).

To **adopt** the *famous dictum* of the *German *military theorist* *Clausewitz about war, law is now the continuation of *politics* by other means. The courts *operate* on a **principle**, not always **acknowledged** but usually present, which *lawyers* call the **principle** of *legality*. It is probably better *described* as a **principle** of **legitimacy**. Some things are *regarded* as ***inherently illegitimate***. For *example*, ***retrospective legislation***, ***oppression*** of *individuals*, ***obstructing*** *access* to a court, acts ***contrary*** to **international law**, and so on.