English Teaching & Learning
https://doi.org/10.1007/s42321-019-00033-2

ORIGINAL PAPER

Academic Spoken Vocabulary in TED Talks: Implications for Academic Listening

TED Talks中的學術口語詞彙:對英語聽力教學之意涵

Chen-Yu Liu[1] · Howard Hao-Jan Chen[1]

Received: 13 April 2019 / Revised: 4 June 2019 / Accepted: 2 July 2019
Published online: 18 July 2019

Abstract

Although TED talks are commonly used as supplementary listening materials in English classrooms, whether they are suitable materials for academic listening is still arguable. This study thus employs the Academic Spoken Word List (ASWL) and the British National Corpus and Corpus of Contemporary American English (BNC/COCA) lists to analyze TED talks' vocabulary profiles in a corpus consisting of transcripts of 2085 such talks from six main topics. The analysis reveals high coverage of the ASWL over TED talks at approximately 90%. The coverage figure is similar to that of the ASWL over academic speech, suggesting that TED talks should be suitable materials for academic listening. Learners are also likely to learn high-frequency academic spoken vocabulary from such talks. This study also discovers that learners can reach the same coverage of TED talks by studying either the 1741 word families in the ASWL or the first 2000 word families in the BNC/COCA lists. The learning load is lower for learners to study the ASWL, thus making it a more suitable vocabulary support for comprehending TED talks. Based on the findings, this study provides several useful suggestions regarding how TED talks can be used in EAP courses.

摘要

儘管TED talks經常被用來當作英語課堂的聽力補充教材,但是它們是否為合適的學術聽力教材仍有爭議。因此,本研究基以學術口語字表(ASWL)和英國國家語料庫及美國當代英語語料庫(BNC / COCA)字表來分析包含六大主題、共2,085個TED talks中的詞彙。分析結果顯示學術口語字表對TED talks有相當高的詞彙覆蓋率(約90%), 亦接近於該字表對學術演講

✉ Chen-Yu Liu
chenyuliu1019@gmail.com

Howard Hao-Jan Chen
hjchen@ntnu.edu.tw

[1] Department of English, National Taiwan Normal University, 162, Section 1, Heping E. Rd, Taipei City 106, Taiwan

的詞彙覆蓋率，顯示TED talks應可做為合適的學術聽力教材，學習者也很可能從TED talks中學習到許多高頻學術詞彙。本研究亦發現不論是學習學術口語字表中的1,741個字族或是BNC / COCA字表中的前2,000個字族，學習者皆可在TED talks中達到相同的詞彙覆蓋率。相較之下，學習學術口語字表的學習負擔較低，因此顯示其較適合做為學生在理解TED talks上的詞彙輔助。本研究亦根據研究結果針對TED talks在學術英文課程中的使用提出教學建議。.

Keywords TED talks · Vocabulary profile · Lexical coverage · Academic listening · Academic spoken vocabulary

關鍵詞 TED talks · 詞彙概述 · 詞彙覆蓋率 · 學術聽力 · 學術口語詞彙

Introduction

With the increasing number of non-native students taking courses in English-language universities or universities in non-English speaking countries adopting English-medium instruction (EMI), a growing number of students need effective academic listening skills [15], because listening competence contributes greatly to academic performance [20]. In order to understand academic speech, good listening ability [22], content knowledge [27], and a large vocabulary size [32] are necessary for a learner. TED talks, presentations freely accessible on the Internet, allow students to enhance their presentation skills [27], improve their academic listening abilities [13], help them understand ideas from a variety of subject areas, and expand their general and academic vocabulary [16]. These talks thus can be considered seriously when teachers think about selecting interesting listening materials for EAP students [9].

TED, which stands for Technology, Entertainment, and Design, is an online database consisting of short English talks delivered by more than 2000 specialists discussing topics in a wide variety of fields, such as technology, culture, global issues, and music. TED talks are commonly used in intensive English programs and university classrooms as supplementary materials because "complex ideas are explained in an engaging and learner-friendly fashion" in these talks ([27], p. 770), thus a good support for learners to more easily understand complex discipline-specific ideas unfamiliar to them. Speakers in TED talks summarize ideas and research findings, provide their views or insights regarding a certain subject, or present creative approach to academic or general issues. Their easy accessibility, wide coverage of topics, and authentic nature have provided rich resources for advanced learners or EAP students. [9, 28].

Despite the usefulness of TED talks for improving students' listening ability, presentation skills, and content knowledge, learners cannot benefit much from them without adequate comprehension, which is strongly correlated with vocabulary size [32, 35]. Therefore, sufficient vocabulary in English is necessary for learners to reach adequate comprehension of TED talks. There have been several studies trying to identify the vocabulary demand of TED talks based on a lexical coverage methodology for providing pedagogical implications for EAP. However, despite their contributions to current understanding of TED talks in terms of their vocabulary profiles, a few limitations should still be addressed, including the limited number of TED talks investigated, and the inappropriateness of adopting written academic word lists to analyze the coverage of academic vocabulary in TED talks, a spoken discourse. There is also a pedagogical need for determining the word lists suitable for vocabulary instruction of TED talks. The present study thus attempts to address these issues by analyzing TED talks' vocabulary profiles based on a large corpus consisting of 2085 TED talks.

Literature Review

Academic Vocabulary and Listening Comprehension

Lexical coverage refers to "the percentage of running words in the text known by the readers" ([24], p. 61). For example, 95% coverage means that 5 out of every 100 words (1 in 20) are unknown words. Lexical coverage is an essential measure because it "allows the calculation of estimates of the vocabulary size necessary for comprehension of written and spoken texts." ([35], p. 457). In terms of spoken discourse, it is suggested that learners should know at least 95% of the words to achieve a high degree of listening comprehension [30]. Previous findings also indicated that 95% lexical coverage is sufficient for adequate listening comprehension [35]. For advanced listening comprehension, however, 98% coverage is needed [32]. The 98% figure is also established aiming at independent learning carried out by learners without support from teachers [18]. Both coverage figures have been consistently used in lexical coverage studies because they "provide a useful indication of whether or not a text may be understood" ([38], p. 1).

Nation [24] estimated the vocabulary sizes necessary for comprehension of different discourse types by calculating the number of word families needed. According to Bauer and Nation [2], a word family consists of "a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately" (p. 253). For example, a word family for *imagine* can include its inflections such as *imagines*, *imagined*, and *imagining*, and its derivatives, such as *imagination*, *imaginative*, and *imaginary.*

Based on Nation's [24] findings, to reach 95% coverage, knowledge of 3000 word families is needed for spoken text and knowledge of 4000 word families is required for written text. To reach 98% coverage, knowledge of 6000–7000 word families is needed for spoken text and knowledge of 8000–9000 word families is required for written text. Schmitt and Schmitt [31] further divided these word families into three bands: high-frequency word families (1st-3000th), mid-frequency word families (3001th–9000th), and the ones after the 9000th word families are categorized as low-frequency. This categorization is pedagogically helpful, as Roche and Harrington [29] suggested that knowledge of high-frequency vocabulary correlated with academic performance and written proficiency.

The concept of word family frequency has been applied to compose discourse-specific word lists, such as academic vocabulary. The Academic Word List (AWL) [6] and Academic Vocabulary List (AVL) [17] are the two predominant word lists that have been widely used for EAP teaching or for material development. They are also used to inform the coverage of academic vocabulary in a variety of academic genres, such as research papers, textbooks, book reviews, and university lectures. [19, 36, 40]. Both word lists were created entirely based on written academic text, with the AWL consisting of 570 word families and the AVL containing 3000 lemmas frequently occurring in academic written discourse. Townsend and Collins [34] have found that the AWL is helpful for improving students' comprehension of academic written texts, and is also useful for developing vocabulary tests and dictionaries [7, 8].

Although the AWL and the AVL provide great help for vocabulary instruction of academic written discourse, they may not be the most useful word lists for improving students' academic spoken vocabulary because academic written word lists may not be able to reflect features of spoken language [9]. A desirable word list for vocabulary instruction of academic spoken discourse had been missing until the recent development of the Academic Spoken Word List (ASWL) by Dang et al. [10], aiming to "help EAP learners from different academic disciplines enhance their comprehension of academic speech" (p. 967). The list was created based on a 13-million-word corpus of academic spoken discourse from a variety of subject areas of four academic speech events (lectures, seminars, labs, and tutorials). The researchers also compared the coverage of the ASWL, the AWL, and the AVL in their academic spoken corpora and revealed that the ASWL had a higher coverage than either the AWL or the AVL did (around 90% compared to roughly 4% and 24%). Their findings suggest that learners might not encounter many words from the AWL or the AVL in academic spoken discourse, while the ASWL words may appear comparatively more frequently. Thus, the ASWL should be more representative of academic spoken discourse than those in the AWL and the AVL, therefore a more suitable word list for instructing academic spoken vocabulary.

Lexical Coverage of TED Talks

Despite TED talks' common usage in language classrooms, there have not been many studies analyzing their vocabulary profiles. A rather small-scaled study was conducted by Nurmukhamedov and Sadler [28] where a 221-word section of a TED talk was analyzed using the General Service List (GSL) [39] and the AWL [6]. Wang [37] compiled an 80,885-word corpus of 10 TED talks up to 20 min in length and compared the vocabulary coverage across the TED talks and the longer lectures in Social Science and Physical Science using Nation's [24] British National Corpus (BNC) list.

Three larger-scaled studies on TED talks' vocabulary profiles are Coxhead and Walls [9], Nurmukhamedov [27], and Wingrove [40]. Coxhead and Walls [9] developed a corpus of 60 six-minute TED talks from six subject areas to analyze their vocabulary loads using the GSL [39], the BNC list [24], and the AWL [6]. They reported that TED talks' vocabulary profile are similar to those of newspapers, novels, and academic texts according to Nation's [24] findings. They also discovered that TED talks' vocabulary profiles are "slightly different from other 'spoken' texts such as movie scripts" (p. 60). The researchers concluded their research by stating that TED talks are closer to written texts than to spoken texts in terms of vocabulary loads. The possible reason for this finding might be due to the carefully scripted nature of TED talks, as the researchers explained.

Nurmukhamedov [27] partially replicated Coxhead and Walls' [9] study by using a more representative corpus of 400 TED talks from four registers (Business, Global issues, Science, and Technology) to identify the vocabulary demand for comprehending TED talks. The researcher suggested that 4000 word families plus proper nouns and marginal words are needed for 95% coverage, and 8000 word families plus proper nouns and marginal words provide 98% coverage. The result allows the researcher to conclude that a large vocabulary size closer to that for written texts is needed to comprehend TED talks.

Aiming at investigating the suitability of TED talks as materials for academic listening, Wingrove [40] compared the AVL [17] representation in a corpus comprised of 60 TED talks and a corpus consisting of 28 lecture series. The comparison revealed that TED talks have significantly lower AVL representation than lecture discourse does. Although some TED talks were found to be similar to lecture discourse in terms of AVL presence, on average, they are not suitably similar to lecture discourse to be used as academic listening materials.

Although previous studies have contributed greatly to our understanding of TED talks from a vocabulary perspective, a few limitations still need to be further addressed. First, the TED talks investigated in previous studies are limited, with 60 TED talks from six topics analyzed by Coxhead and Walls [9], 49 TED talks from three topics by Wingrove [40], and 400 TED talks from four topics by Nurmukhamedov [27]. There have been over 2000 TED talks available on its website. The analysis of TED talks' vocabulary profiles could be more representative if a larger number of TED talks is analyzed.

Second, Wingrove [40] questioned the suitability of TED talks as academic listening materials for a lower AVL representation was found in TED talks than in lecture discourse. However, the AVL was developed based on academic written text. Given the substantial differences of vocabulary use between academic speaking and writing [33], it may not be the most appropriate approach to analyze the academic vocabulary coverage in academic spoken texts using a list of frequent academic written vocabulary. Moreover, some everyday spoken language should appear in academic spoken discourses, which cannot be reflected in academic written word lists [9]. Thus, using the AWL or AVL, both created based on academic written discourse, for examining the coverage of academic vocabulary in TED talks may not yield the fairest result of comparison. The authors thus suggest the ASWL, a word list developed based on academic spoken text, be more appropriate for examining the lexical coverage of academic vocabulary in TED talks and determining whether they are suitable materials for academic listening.

Third, from an EAP instruction perspective, it is well established that there is a strong correlation between vocabulary size and listening comprehension [32, 35]. Although the AWL is used widely as a guidance for academic vocabulary in EAP courses, it in fact has a much lower coverage over academic spoken discourse than written one [27]. Coxhead and Walls' [9] and Nurmukhamedov's [27] studies found that the coverage of the AWL over TED talks ranges from 3.79 to 3.90%, and 4.41% coverage of the AWL over academic lectures was reported by Dang and Webb [11]. Evidently, learners benefit limitedly from learning the AWL words for improving their vocabulary knowledge of academic spoken discourse. Thus, it is pedagogically helpful to analyze TED talks' vocabulary profiles and identify a word list that can serve as a suitable and efficient vocabulary guidance for teachers to help learners improve their vocabulary knowledge and reach better comprehension of TED talks.

With the aim to fill the gaps from previous research, this study has drawn on a corpus consisting of 2085 TED talks of six main topics delivered from 1984 to 2016, and analyzed the presence of high-frequency academic spoken words in the corpus.

This study also analyzed TED talks' vocabulary profiles based on a general word list, the BNC/COCA lists [25] and the ASWL [10], to determine suitable vocabulary guidance for better comprehension of TED talks. The vocabulary profiles of the six main topics of TED talks were also compared to see whether vocabulary demand varies across topics. Two research questions were proposed to guide the current study:

1. What are TED talks' vocabulary profiles based on the Academic Spoken Word List and BNC/COCA lists?
2. Do the six main topics in the TED corpus have different vocabulary profiles?

Method

The TED Corpus

The talks of the TED corpus were gathered using the Web Inventory of Transcribed and Translated Talks tool (WIT^3; [4]), which retrieved their video and audio transcripts automatically from the TED Talks website. The information regarding the talks, including their titles, subtitles, speaker's names, and dates were also retrieved using the tool. The talks gathered by the tool were used to compile the TED corpus, which consists of 2085 TED talks. These talks were delivered by 1786 different speakers between 1984 and 2016, with 9 talks delivered before the year 2000, 869 talks given between 2000 and 2010, and 1207 talks delivered between 2011 and 2016. In terms of the length, 245 talks are below 6 min, 511 talks range from 6 to 12 min, 842 talks range from 12 to 18 min, and 487 talks are over 18 min. The size of the TED corpus is about 4.37 million words.

The talks in the TED corpus were further categorized into six topics for investigation of their individual vocabulary profiles. It should be noted that categorization of all the current TED talks was a very challenging process because they cover "almost all topics", as stated on the TED Talks website. Each TED talk on the website has a series of topic tags, with most talks belonging to multiple topical categories. To decide what the main topics are, each talk's topic tags were retrieved from the website and were used to determine the relatively common topics. Finally, the top six topics were identified, which are Culture, Design, Entertainment, Global Issues, Science, and Technology. Table 1 summarizes the numbers of talks and word counts of each of the six topics of TED talks.

A problem with categorization of all the talks in the corpus into topics was that the number of talks and the word counts of each topical category were not balanced.

Table 1 Numbers of talks and word counts in the TED corpus, by topic

Topic	Number of talks	Word count
Culture	330	738,426
Design	260	484,779
Entertainment	219	386,754
Global Issues	354	756,028
Science	308	673,305
Technology	614	1,331,357
Total	2085	4,370,649

However, since the proportion of each topic on the TED Talks website is not originally balanced, and great variations of the length also exist across talks, it is thus not surprising to see uneven numbers of talks and word counts across different topical categories. The researchers acknowledged the problem of the unbalanced sizes of the six topical sub-corpora, but this was unpreventable if an extensive investigation of the TED talks' vocabulary profiles were to be conducted.

Data Analysis

Two word lists, the ASWL [10] and the BNC/COCA lists [25], were used to analyze the vocabulary profiles of the TED talks in the current corpus.

The ASWL contains 1741 word families, created from a 13-million-word corpus comprising of a variety of academic spoken discourses. Based on frequency, the 1741 word families were further divided into four levels, with level 1 containing 830 word families, level 2 consisting of 456 word families, 380 word families in level 3, and 75 word families in level 4. This list was created solely based on academic spoken discourse, thus a more appropriate word list for analyzing TED talks, a spoken discourse, than other word lists that were compiled based on written discourse, such as the AWL and the AVL.

The BNC/COCA lists were also used to profile the current corpus, which were compiled by Nation [25], containing the general 25,000 word families based on the BNC and COCA. They were further divided into 25 sub-lists with each containing 1000 word families. Four additional lists were also created to include proper nouns, marginal words, transparent compounds, and acronyms respectively.

Considered the strong written nature of the BNC, the BNC/COCA 2000, which contains the first 2000 word families in the BNC/COCA lists, was created based on a balanced corpus of written and spoken English, specifically to help material development for second language learners [25]. It is also considered the most suitable high-frequency word list for second language learners by Dang and Webb [12] based on their analyses after comparing the BNC/COCA 2000 with the BNC 2000 list [23], the GSL [39], and the new GSL [3] in terms of teachers' perceptions of vocabulary usefulness, lexical coverage, and learner vocabulary knowledge. The BNC/COCA 2000, a general high-frequency word list, was thus used to compare with the ASWL, an academic high-frequency word list, to see which one may be a more suitable word list as vocabulary guidance for teachers to improve learners' comprehension of TED talks.

The AntWordProfiler [1] was used to run the ASWL and the BNC/COCA lists over the TED corpus because it allows users to upload their own word family lists for vocabulary profiling and is deemed the most appropriate tool to run the BNC/COCA lists for coverage research [25]. The researchers thus used it to analyze the vocabulary profiles of the whole TED corpus and of the six sub-corpora of the six topics based on the AWSL and BNC/COCA lists.

Results and Discussion

Academic Spoken Word List Representation in the TED Corpus

Overall, the ASWL covered 89.6% of the current TED corpus, with levels 1, 2, 3, and 4 covering 82.5%, 4.7%, 2.2%, and 0.2% of the TED corpus respectively. Dang et al. [10] also reported similar coverage (90.13%) of the ASWL over their academic spoken corpus, which included four speech events: lectures, seminars, labs, and tutorials. This showed that the current TED corpus and Dang et al.'s academic spoken corpus have comparable coverage by the ASWL. In other words, students will encounter the ASWL words in TED talks as frequently as they will do in academic spoken discourse. Language learners are highly likely to learn academic spoken vocabulary from TED talks.

Compared to the high coverage figure of the ASWL over the current TED corpus, the coverage figures of the AWL or the AVL over TED talks were evidently low. Coxhead and Walls [9], and Nurmukhamedov [27] reported low coverage of the AWL over TED talks, ranging from 3.79 to 3.90%. Wingrove [40] reported that the core AVL representation in his corpus of TED talks ranged from 3.70 to 5.70%, and the representation figures of the wide AVL were between 1.80 and 2.45%. It was not surprising to see the low coverage of TED talks by either the AWL or the AVL since they were both created entirely based on academic written data. These findings revealed the variation of vocabulary use between academic spoken and written discourses, emphasizing the importance of choosing proper word lists for vocabulary instruction of certain discourse type.

As Webb and Nation [38] explained, coverage figures are important for language teachers and learners for they "provide a useful indication of whether or not a text may be understood" (p. 1). These figures can also help teachers to determine which word lists may be more suitable to be used for improving students' vocabulary as well as their comprehension of certain discourse types efficiently.

Among the ASWL, the AWL, and the AVL, we considered the ASWL more suitable for vocabulary instruction of TED talks. The ASWL contains many more word families than the AWL (1741 vs. 570), so it is not surprising that the ASWL has a higher coverage over TED talks compared to the AWL (89.6% vs. around 4%). However, if we examine the learning load, studying the ASWL may be more efficient than studying the AWL for improving learners' vocabulary and comprehension of TED talks. If a learner studies all the 570 word families in the AWL, s/he would supposedly know around 4% of the words in TED talks. On the other hand, even when a learner only studies the 830 word families in the level 1 sublist of the ASWL, which only has 260 more word families than the AWL, potentially s/he would achieve around 83% vocabulary coverage of TED talks. Apparently, it would be more efficient for learners to study the ASWL rather than the AWL if the goal is to improve their vocabulary and comprehension of TED talks.

Similarly, although the AVL contains a larger number of word families compared to the ASWL (1983 vs. 1741), its coverage over TED talks is actually much lower than the ASWL's (around 5% vs. 89.6%). Studying the AWSL is thus more efficient for learners because the learning load is lower but a much higher coverage could be obtained.

The higher coverage of the TED talks in the current corpus by the ASWL than the AWL and the AVL [9, 27, 40] showed that TED talks' vocabulary profiles are closer to that of academic spoken discourse than that of academic written discourse. We thus suggest that in terms of improving learners' vocabulary and comprehension of TED talks, using a word list that has a higher coverage over TED talks and shares the spoken nature as the vocabulary guidance could be more helpful and efficient. Since the ASWL has a relatively high coverage over TED talks compared to either the AWL or the AVL, and can also reduce the learning load, it should be a more suitable vocabulary guidance for TED talks.

Despite the wide use of the AWL and the AVL in EAP courses, they may not be sufficient for improving learners' academic listening, considering their low coverage over TED talks and other academic spoken discourses [11]. Language learners may have little chance of learning AWL or AVL words from TED talks or other academic spoken discourses [27]. This study thus suggests that the AWL and the AVL be used for improving learners' comprehension of written text, while the ASWL be used to expand learners' vocabulary for academic listening.

To explore whether the six sub-corpora of the six topical categories in the TED corpus have similar or different vocabulary profiles, this study also analyzed their coverage by the ASWL. Table 2 summarizes the coverage of each of the six sub-corpora by the level 1 to 4 word families in the ASWL.

In terms of the coverage by all the word families in the ASWL, the six topics have similar coverage figures, ranging from 88.53 to 89.98%. Figure 1 presents the distribution of the four-level word family lists of the ASWL in the six sub-corpora.

Dang et al. [10] indicated that levels 1 and 2 of the ASWL contain general high-frequency words based on the most frequent 2000 BNC/COCA word families [25]. Across the six topical categories, Entertainment had the highest coverage by levels 1 and 2 of the ASWL, suggesting that language learners are more likely to encounter general high-frequency words in the talks of this topical category. This makes the talks categorized as Entertainment more suitable materials for beginning EAP learners. Such talks can also be considered a starting point for them to see what an academic speech is like.

Level 3 and level 4 word families of the ASWL are academic words frequently occurring in academic speech but fall outside general high-frequency words [10]. Technology and Science have similar coverage figures of level 3 and 4 word families, 2.64% and 2.65% respectively, both of which were higher than those of the other four topical categories. This suggests that more high-frequency academic spoken words appear more often in the talks related to Technology and Science.

Some pedagogical implications can be drawn from these results. When selecting materials from the TED Talks website for EAP beginning learners, teachers can start with talks belonging to the topical categories which have higher coverage of the level 1 and 2 word families of the ASWL, such as those in Entertainment. Talks of the topical categories containing more frequent academic spoken words, such as those in Science and Technology, may be used for advanced EAP learners or as challenging materials.

Table 2 Coverage (by percentage) of the six sub-corpora of the six topical categories by the ASWL

Topic	Culture	Design	Entertain	Global	Science	Tech
Level 1	83.13	82.71	83.54	82.28	81.65	82.43
Level 2	4.49	4.66	3.44	4.99	4.86	4.91
Level 3	1.99	2.07	1.43	2.43	2.46	2.43
Level 4	0.1	0.16	0.12	0.1	0.19	0.21
Total	89.71	89.6	88.53	89.8	89.16	89.98

Profiling TED Talks with the BNC/COCA Lists

The coverage of the BNC/COCA lists over the TED corpus were also analyzed to find out how many word families plus the additional words (proper nouns, marginal words, transparent compounds, and acronyms) are needed for language learners to reach adequate listening comprehension of TED talks. Note that transparent compounds and acronyms were also counted separately, in addition to proper nouns and marginal words which are usually included in separate lists in lexical coverage studies [9, 27]. Nation [26] indicated that transparent compound words can be easily understood by learners once their parts are known, with little or no learning required. “If they (transparent compounds) are counted as new unknown words, they would inflate the number of unknown words” ([26], p. 68). Similarly, acronyms are comprised of already known parts and they are usually “explicitly related to their parts on their first occurrence” for receptive purposes ([26], p. 85). Therefore, separate lists of transparent compound words and acronyms were added to the current analysis to avoid inflating the amount of vocabulary necessary for comprehending TED talks.

The BNC/COCA 2000, the first 2000 high-frequency general word families, has a coverage of 91.71% over the TED corpus plus the additional words. The 95% coverage is reached at around 3000 word families (93.72%) plus the additional words (1.99%). To reach 98% coverage of the current TED corpus, around 6000 word families (96.2%) plus the additional words (1.99%) are required.

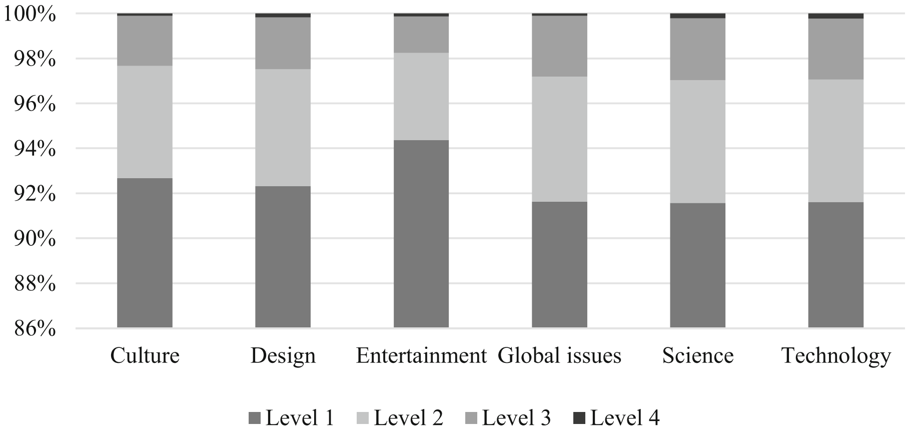

Fig. 1 Distribution of the four-level word family lists of the ASWL in the six sub-corpora

Nurmukhamedov [27] reported a larger vocabulary demand for good comprehension of TED talks, with 4000 word families required for 95% coverage and 8000 word families needed for 98% coverage plus proper nouns and marginal words. Despite the different vocabulary demands reported by Nurmukhamedov [27] and the present study, without including the additional lists, Nurmukhamedov [27] and the present study both reported similar coverage figures of the first 3000 word families over TED talks (93.35% and 93.72% respectively). The reason for the discrepancy may be due to the fewer additional lists utilized in Nurmukhamedov's [27] analysis. This study used four additional lists (proper nouns, marginal words, transparent compounds, and acronyms) in addition to the main BNC/COCA lists to analyze TED talks' vocabulary profiles, resulting in 1.99% coverage by the four additional lists. However, Nurmukhamedov [27] only included two additional lists (proper nouns and marginal words) and identified 0.98% coverage by those two lists. Some specialized words might not be identified with the two additional lists used by Nurmukhamedov [27], thus resulting in a larger vocabulary demand. The comparison revealed that a smaller vocabulary demand may be required for good comprehension of TED talks and highlighted the importance of teaching learners the words in the additional lists [14, 21], since many of them occur in TED talks.

To reach adequate and/or advanced comprehension of academic lectures, Dang and Webb [11] suggest that students need to know 4000 to 8000 word families. TED talks can be suitable preparatory materials for EAP learners for their vocabulary demand is slightly lower but mostly overlapped (3000–6000 word families) with that of academic lectures. EAP teachers can thus use TED talks as materials to help learners get ready for academic lectures and bridge the vocabulary gap.

To further analyze the discrepancies across the vocabulary profiles of the six topical categories in the TED corpus, the vocabulary demands for 95% and 98% coverage of the six topical categories were analyzed, as summarized in Table 3.

To reach 95% coverage, 3000 word families plus the additional words are required for all of the six topical categories. However, the numbers of word families needed for 98% coverage are different across topics. To reach 98%, Global issues require 5000 word families plus words in the additional lists; and Culture, Design, Entertainment, and Technology need 6000 word families plus words in the additional lists. Science

Table 3 Percentage of the BNC/COCA lists plus the additional words required to reach 95% and 98% coverage over each topic

Word list	Culture	Design	Entertainment	Global issues	Science	Technology
3000	96.07[a]	95.79[a]	95.58[a]	96.22[a]	95.06[a]	95.56[a]
4000	97.18	96.98	96.78	97.32	96.5	96.86
5000	97.92	97.81	97.64	98.08[b]	97.42	97.67
6000	98.35[b]	98.21[b]	98.12[b]	98.43	97.98	98.12[b]
7000	98.64	98.52	98.39	98.72	98.37[b]	98.44

[a] Reaching 95% coverage

[b] Reaching 98% coverage

requires the highest number of word families than the other five topics to reach 98% coverage, with 7000 word families plus words in the additional lists required.

These results showed that some topics of TED talks can be easier for second language learners in terms of vocabulary demand. Global issues require the least word families to reach 98% coverage, which could be considered materials more suitable for beginning EAP learners. A lot more word families are needed to reach 98% for topics of Science than the other five topics, suggesting that talks of Science-related topics be introduced to EAP learners with larger vocabulary size.

Coxhead and Walls [9] reported that 5000 word families plus proper nouns were needed for 95% coverage for all the six TED talk topics in their corpus. The larger vocabulary demand than that found by this study may likewise be due to the use of fewer additional lists in their study than the present study, as also explained previously. If specialized words in TED talks are identified by more additional lists, the vocabulary demand for TED talks could be more precisely measured. Koveleva (2012) and Dang et al. [10] have emphasized the importance of teaching proper nouns explicitly for better listening comprehension. Seeing many words in TED talks are actually from the additional lists; this study suggests that, in addition to proper nouns, the words from the other three additional lists including acronyms, transparent compounds and marginal words should also be taught explicitly for learners' better comprehension.

One of the purposes of this study is to identify a suitable word list for vocabulary guidance for TED talks. The comparison of the coverage of the TED corpus by the ASWL and the BNC/COCA 2000 revealed that the ASWL should be a more suitable vocabulary guidance for improving learners' comprehension of TED talks especially for EAP learners. If a learner studies all the 1741 word families in the ASWL, he/she may reach 90% coverage of TED talks. If the additional words (proper nouns, marginal words, transparent compounds, and acronyms) are known, the potential coverage for him/her is 92%. Although these figures are similar to that provided by the BNC/COCA 2000 (91.71% coverage with additional words included), note that the learning load is lower for learners to learn the 1741 word families in the ASWL than the 2000 word families in the BNC/COCA 2000 in order to reach the same coverage. Thus, it is suggested that ASWL be used as the guidance for vocabulary instruction of TED talks, especially for learners who have not mastered the most frequent 2000 word families and whose learning goal is EAP. As for learners who have mastered the most frequent 2000 word families when they attend EAP courses, they only need to learn the 455 word families in the ASWL that are beyond the first 2000 word families in the BNC/COCA lists to reach over 90% coverage [10].

Pedagogical Implications

The results of this study showed that TED talks and academic spoken discourse, such as lectures and seminars, similarly have high coverage figures by the ASWL, suggesting that learners can encounter high-frequency academic spoken words frequently in both discourse types. This allows the researchers to suggest that TED talks be suitable materials for improving academic listening especially for language learners who wish to attend academic lectures delivered in English.

Profiling TED talks with the ASWL and the BNC/COCA lists has provided useful suggestions for teaching English for academic purposes. For EAP teachers, TED talks are suitable materials for expanding learners' academic spoken vocabulary because of the frequent presence of the ASWL words in these talks. Also, the vocabulary demand for TED talks are slightly lower than academic lectures, thus making them especially useful for preparing beginning EAP learners for academic speech. As 90–95% coverage of academic spoken discourse is considered an important goal for EAP leaners [10], a high coverage of the ASWL over TED talks (around 90%), identified by the present study, allows the researchers to suggest the ASWL be the suitable vocabulary guidance for TED talks.

About 2% words in TED talks are proper nouns, marginal words, transparent compounds, or acronyms. Cobb [5] indicated that "proper nouns are not lexical items" (p. 187) so they do not need to be taught. However, Koveleva (2012) and Dang et al. [10] emphasized that proper nouns should be explicitly taught or pretaught because they may carry some connotations in addition to the meaning, which may facilitate learners' comprehension [14]. In addition to proper nouns, the authors also argue that marginal words, transparent compounds, or acronyms should also be taught explicitly or pretaught to learners. Teachers can identify such words from a TED talk and preteach them before learners watch it, which can help reduce the talk's vocabulary load so learners can focus more on understanding the content without being distracted by these words.

To enhance learners' topic-related knowledge, teachers can select a series of TED talks centering on similar topics and preteach some important and topic-related words from the selected talks. This not only can reduce the vocabulary load of the talks, but also can draw learners' attention to the target words teachers wish them to learn. This approach is also believed to enhance learners' technical or topical vocabulary [27].

Conclusion

This study adopted a spoken academic high-frequency word list, the ASWL, to identify the coverage of academic vocabulary in TED talks. The analysis showed a high coverage of the current TED corpus by the ASWL, at around 90%, which is similar to the coverage of academic spoken discourse by the ASWL [10]. These findings suggest that learners are very likely to learn the ASWL words from TED talks, thus making them suitable materials for expanding learners' academic spoken vocabulary and improving their academic listening.

Profiling the TED corpus with the ASWL and the BNC/COCA lists showed that similar coverage (92%) of the TED corpus can be reached by all the 1741 word families in the ASWL or by the 2000 word families of the BNC/COCA 2000 plus words in the additional lists. Learners can study fewer word families to reach the same coverage with the help of the ASWL, which is thus considered a more suitable vocabulary guidance for improving learners' comprehension of TED talks than the BNC/COCA 2000.

This study also revealed that 95% coverage of TED talks is reached at 3000 word families, and 6000 word families are needed to reach 98% coverage. This finding suggests that in addition to high-frequency vocabulary, learners will also need to know mid-frequency vocabulary words in order to reach better comprehension of TED talks.

The analysis of the six topical categories of the TED corpus also revealed that there are variations across the six topics of TED talks based on their vocabulary profiles. Relatively high coverage of the ASWL was found in talks related to Technology while those in Entertainment have comparatively low ASWL coverage. Talks about Science were found to require a larger vocabulary size for good comprehension than the other topics while those associated with Global issues require the least word families to reach adequate comprehension. These findings provide useful pedagogical suggestions for teachers when considering which TED talks to select for their students of different vocabulary levels.

There are several limitations that this study failed to address and further investigations are thus required. The sizes of the six topical sub-corpora were not well-balanced. Yet, the distribution of the talk topics was difficult to be controlled if TED talks were to be analyzed extensively. It is suggested that future studies be aware of whether the sizes of sub-corpora are comparable. Also, this study focused on identifying the vocabulary demands of the six topics of TED talks to inform language teachers how many word families are needed for their students to comprehend such talks. It would also be interesting to explore the vocabulary items that are specific to certain topical categories. Such vocabulary words can be technical vocabulary for certain subject areas, which can be further taught or emphasized to students in related fields to enhance their disciplinary vocabulary knowledge. This issue thus should be further explored in future research so students can benefit more from TED talks. Certain TED talks might even be especially helpful to prepare students for speech in certain disciplinary fields.

With more and more TED talks conferences organized and held, there will surely be more materials available on the TED Talks website. More studies are still needed to determine the ones suitable for students of different proficiency levels and of different learning needs. In sum, although this study focuses on promoting the importance of using TED talks with the ASWL to improve learners' academic listening and academic spoken vocabulary, it should be noted that knowledge of academic written and spoken vocabulary words are both essential for learners to achieve academic success.

Acknowledgements This work was financially supported by the "Chinese Language and Technology Center" of National Taiwan Normal University (NTNU) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

References

1. Anthony, L. (2014). *AntWordProfiler (version 1.4.1) [computer software]*. Tokyo: Waseda University. Retrieved from http://www.laurenceanthony.net/software . Accessed 5 Mar 2019.
2. Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*, 253–279.
3. Brezina, V., & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics, 19*, 1–28. https://doi.org/10.1075/ijcl.19.1.01bre.
4. Cettolo, M., Girardi, C., & Federico, M. (2012). *Wit3: Web inventory of transcribed and translated talks.* In M. Cettolo, M. Federico, L. Specia & A. Way (Eds.), *Proceedings of the 16th Annual Conference of*

the European Association for Machine Translation (EAMT 2012) (pp. 261–268). Retrieved from http://hltshare.fbk.eu/EAMT2012/html/Papers/59.pdf . Accessed 27 Dec 2018.

5. Cobb, T. (2010). Learning about language and learners from computer programs. *Reading in a Foreign Language, 22*, 181–200.
6. Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213–238. https://doi.org/10.2307/3587951.
7. Coxhead, A. (2011). The academic word list 10 years on: research and teaching implications. *TESOL Quarterly, 45*, 355–362. https://doi.org/10.5054/tq.2011.254528.
8. Coxhead, A. (2016). Reflecting on Coxhead (2000) "a new academic word list". *TESOL Quarterly, 50*, 181–185. https://doi.org/10.1002/tesq.287.
9. Coxhead, A., & Walls, R. (2012). TED talks, vocabulary, and listening for EAP. *TESOL ANZ Journal, 20*, 55–65.
10. Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning, 67*(4), 959–997.
11. Dang, T. N. Y., & Webb, S. (2014). *The lexical profile of academic spoken English* (Vol. 33, pp. 66–76). English for Academic Purposes. https://doi.org/10.1016/j.esp.2013.08.001.
12. Dang, T. N. Y., & Webb, S. (2016). Evaluating lists of high-frequency words. *ITL-International Journal of Applied Linguistics, 167*, 132–158.
13. De Chazal, E. (2014). Prepare English language students for academic listening. British Council. Britishcouncil.org. Retrieved from https://www.britishcouncil.org/voices-magazine/prepare-english-language-students-academic-listening . Accessed 9 Apr 2019.
14. Erten, I., & Razi, S. (2009). The effects of cultural familiarity on reading comprehension. *Reading in a Foreign Language, 21*, 60–77.
15. Flowerdew, J., & Miller, L. (1997). The teaching of academic listening comprehension and the question of authenticity. *English for Specific Purposes, 16*(1), 27–46.
16. Floyd, M., & Jeschull, L. (2012). Teaching with TED talks: authentic and motivational language instruction. *Newsletter of the Video and Digital Medial Interest Section*. Retrieved from newsmanager.commpartners.com/tesolvdmis/issues/2012-08-10/ 12.html. Accessed 19 Mar 2019.
17. Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327. https://doi.org/10.1093/applin/amt015.
18. Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.
19. Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly, 41*(2), 235–253.
20. Jeon, J. (2007). *A study of listening comprehension of academic lectures within the construction-integration model* (unpublished doctoral dissertation), The Ohio State University, IL.
21. Kobeleva, P. (2012). Second language listening and unfamiliar proper names: Comprehension barrier? *RELC Journal, 43*, 83–98. https://doi.org/10.1177/0033688212440637.
22. Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes, 10*, 79–88. https://doi.org/10.1016/j.jeap.2011.03.001.
23. Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 3–14). Amsterdam: John Benjamins.
24. Nation, I. S. P. (2006). How large a vocabulary is needed to reading and listening? *Canadian Modern Language Review, 63*, 59–82. https://doi.org/10.3138/cmlr.63.1.59.
25. Nation, I. S. P. (2012). The BNC/COCA word family lists. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation. Accessed 9 Apr 2019.
26. Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins.
27. Nurmukhamedov, U. (2017). Lexical coverage of TED talks: implications for vocabulary instruction. *TESOL Journal, 8*(4), 768–790.
28. Nurmukhamedov, U., & Sadler, R. (2011). Podcasts in four categories: applications to language learning. In B. Facer & M. Abdous (Eds.), *Academic podcasting and mobile assisted language learning* (pp. 176–195). Portland: Book News.
29. Roche, T., & Harrington, M. (2013). Recognition vocabulary knowledge as a predictor of academic performance in an English as a foreign language setting. *Language Testing in Asia, 3*(1), 12.
30. Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching, 50*(2), 212–226.

31. Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484–503.
32. Stæhr, L. T. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition, 31*, 577–607. https://doi.org/10.1017/s0272263109990039.
33. Thompson, P. (2006). A corpus perspective on the lexis of lectures, with a focus on economics lectures. In K. Hyland & M. Bondi (Eds.), *Academic discourse across disciplines* (pp. 253–270). Frankfort: Peter Lang.
34. Townsend, D., & Collins, P. (2009). Academic vocabulary and middle school English learners: An intervention study. *Reading and Writing, 22*(9), 993–1019.
35. van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: the same or different from reading comprehension? *Applied Linguistics, 34*, 457–479. https://doi.org/10.1093/applin/ams074.
36. Vongpumivitch, V., Huang, J., & Chang, Y. (2009). Frequency analysis of the words in the academic word list (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes, 28*(1), 33–41.
37. Wang, Y. (2012). An exploration of vocabulary knowledge in English short talks: a corpus driven approach. *International Journal of English Linguistics, 2*, 33–43.
38. Webb, S., & Nation, P. (2013). Computer-assisted vocabulary load analysis. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–10). Malden: Wiley-Blackwell.
39. West, M. (1953). *A general service list of English words*. London: Longman, Green and Co.
40. Wingrove, P. (2017). How suitable are TED talks for academic listening? *Journal of English for Academic Purposes, 30*, 79–95.