European Journal of Foreign Language Teaching
ISSN: 2537 - 1754
ISSN-L: 2537 - 1754
Available on-line at: www.oapub.org/edu

doi: 10.46827/ejfl.v5i3.3577 Volume 5 | Issue 3 | 2021

A CORPUS-BASED STUDY ON THE EDUCATION-RELATED TED TALKS BETWEEN NATIVE AND NON-NATIVE SPEAKERS

Hieu Manh Do[i]
Department of English Language,
Hong Bang International University,
Vietnam

Abstract:
This study aims to find out the frequency word lists in the TED talks in the education field as well as the comparison of the language used by native speakers (NS) and non-native speakers (NNS). The researcher collected four transcripts (two from NS and the others two from NNS) from the TED talks. AntConc is the main software that would be used to investigate the frequency word lists. Data collection includes two steps: (1) collecting the four transcripts of TED talks and (2) listing top 10, 20, and 100 frequency word lists of TED talks corpus of NS and NNS, separately. The findings found that both speakers usually use functional words more than content words. However, content words play a pivotal role in making a full meaning sentence.

Keywords: TED talks, AntConc, frequency word lists

1. Introduction

A corpus is "*a collection of authentic language either written or spoken, which has been complied for a particular purpose*" (Flowerdew, 2011, p.3) (as cited in Sinclair, 1991; Stubbs, 1996; Biber et al., 1998; Hunston, 2002). As for the purpose of linguistics, corpus is considered as authentic teaching and learning ways in the field of second or foreign language acquisition (Wang, 2012). In other words, learners could learn language in the real and rich material resource under this framework (McCarthy, 2001; Reppen, 2009).

Regarding spoken texts, language corpora could have capability to "*add multimedia elements, such as video clips, to corpora of spoken language*" (O'keeffe, McCarthy & Carter, 2007, p.2). TED talks is selected to build the TED corpus that could be used in teaching and learning language. In order to find out the differences of language uses between native and non-native English speakers in spoken texts, this study first aims to explore the frequency word lists in TED talks in education topic. The second purpose is to

[i] Correspondence: email domanhhieubc@gmail.com

compare the language used by native and non-native speakers according to the frequency word lists analysis. Thus, the main structure of this study contains two sections. The first section analyzed the frequency word lists in TED talks of native speakers and non-native speakers, separately. The second part categorized the differences in the language used by native and non-native speakers when they made a speech. It is hoped that the findings might help students, especially, EFL Vietnamese learners understand how English language is used and delivered in the real contexts between native and non-native English speakers.

2. Literature Review

2.1 Definition of Terms

2.1.1 TED Talks

TED talks (www.ted.com) is an abbreviation of Technology, Entertainment, and Design. TED started in 1984 as a nonprofit devoted to spreading ideas. Every talk is about 18 minutes or less so it can be considered as a short talk with fewer than 20-minute presentation (Loan, 1990). All the topics are from specific fields such as business, global issues, science in many different languages, and people from every discipline and culture.

2.1.2 AntConc

AntConc was created by Laurence Anthony from Waseda University. This program is helpful for analyzing electronic texts in order to find and reveal patterns in language. The users need to type or copy and paste texts, and all files need to be stored into plain-text files (txt) then saved. *AntConc* program contains seven tools with specific functions: word list, concordance, collocates, clusters, N-gram, plot, and keyword list.

2.1.3 Frequency Word Lists

Frequency word lists are lists of words that appear in the text many times and on the top in the word list, followed by less frequent words.

2.2 Studies Related to TED Talks Corpus

According to the literature of the effective TED talks corpus in second language learning and teaching, several studies have been investigated to explore the linguistics functions (Chen I-Chieh, 2017; D'avanzo, 2015; Wang, 2012). Table 1 below presents the titles of these studies, followed by the review in detail and research questions.

Table 1: Several Previous Studies to TED Talks Corpus

Authors	Studies	Year
Wang	An Exploration of Vocabulary Knowledge in English Short Talks – A Corpus Driven Approach.	2012
D'avanzo	Speaker Identity vs. Speaker Diversity: The Case of TED Talks corpus.	2015
Chen I-Chieh	A Corpus-based Study of the Business-related TED Talks from 2006 to 2016.	2017

In Wang's (2012) study, three corpus tools (*AntConc*, *RANGE*, and *KfNgram*) were applied to explore the frequency word lists, concordance lines, vocabulary coverage as well as

lists of lexical bundles. Wang collected data from the TED talks and the British Academic Spoken English (BASE) to compare the occurrence of words in the TED talks and the BASE. According to this goal, the researcher chose 80 lectures and 20 seminars of physical science and social science themes from the BASE corpus and 30 TED talks videos from science, technology, global issues, and business topics. The results indicate that the most frequently used words in the TED talks corpus and the BASE corpus were not different. High-frequency words appeared on the top ten of both corpus were functional words such as *of, the,* but the order was not the same: "of" appeared as top three in the TED talks corpus but ranked as top two in the BASE corpus. In addition, the results of concordance lines and lexical bundles presented frequency words in the beginning, middle, and ending part of English short talks.

In Taiwan, Chen conducted a corpus-based study of the business-related TED talks from 2006 to 2016 (2017). On the one hand, the researcher explored the frequency word lists of 316 TED talks videos. On the other hand, she wanted to analyze the lexical coverage of TED talks corpus using the TOEIC word family list and understand the function of watching the TED talks to prepare the TOEIC test by interviewing three Air Force officers. Therefore, the primary methods employed for this study were the combination of quantitative and qualitative approach. In this study, Chen used different software from Wang's study, *AntConc* and *AntWordProfiler* for analysis. The researcher carried out three main steps for data collecting: collecting 316 transcripts of the TED talks, establishing the TOEIC word family list, and conducting a semi-structured interview. From this procedure, the researcher revealed that among 100 frequency words of the TED talks, functional words occupied 68 percent and 32 percent content words. This result was the same as Wang's (2012) study when functional words appeared more frequently on the top in the word lists, although they are "not informative" than content words. As Wang mentioned, however, functional words become more important and essential when combined with other words such as *look, part*, for example, *look at the, part of the*. Thus, functional words are essential and connected to other words in order to create sentences. Besides, the researcher found that the lexical coverage of TED talks corpus's lexical coverage in the TOEIC word family list is 13.02%. Finally, three participants recognized that their TOEIC scores improved after watching the TED talks video and considered TED talks as a useful tool for people who wanted to improve their English skills, especially, reading and listening.

Another researcher, D'avanzo (2015) also conducted a TED corpus with 1,131 talks (2.628.455 tokens). However, the researcher carried out with different purposes, compared with the two aforementioned studies. D'avanzo (2015) focused on analyzing speakers who came from different fields. He investigated the concepts of "speaker identity" and "speaker diversity" in the TED corpus to see the differences between speakers. The researcher focused on the different rhetorical choices made by speakers with different professional categories. These included: academics – all people working at universities, schools, research centers, professionals, doctors, politicians, artists, literary writers, lay (VIP), lay (ordinary people). As a result, the researcher tried to explore the

frequency of reader pronouns (you and your) concerning a distinction between experts and lay speakers. The results indicate that the experts used reader pronouns more frequently than non-specialists did. The researcher also wanted to see the frequency of obligation modals (have to) regarding a distinction between experts and lay speakers. The researcher concluded that the number of directives in the whole corpus used by the laymen was not high.

In conclusion, according to the previous research, all results proved the significant role of corpus linguistics in teaching and learning English with different purposes. Moreover, teachers and students can use different software to analyze different language aspects from the TED talks corpora such as written and spoken languages. In order to bridge the research gaps, therefore, the current study investigates the frequency word lists in the TED talks among native and non-native speakers in education field, which has not been conducted by the previous studies. Then the different language uses between NA and NNS would be sought out. The three research questions guided this study:

1) What is the frequency word list in TED talks of native speakers in education topic?
2) What is the frequency word list in TED talks of non-native speakers in education topic?
3) Are there any differences in the frequency word lists between native and non-native speaker?

3. Methodology

3.1 Data Collection

In this corpus-based study, the researcher analyzed four TED talks videos: two from native speakers and two from non-native speakers. The researcher used the *AntConc* program to analyze the frequency word lists. This study aims to analyze the differences in the language used by NS and NNS. Thus, the researcher tried to compare videos with almost the same word numbers in order to avoid influencing the result. Although the word numbers of all videos between NS and NNS are not the same, the distance of total word numbers between NS and NNS videos is almost close so it can be acceptable (6323-word numbers of NS and NNS-6135). The topics of the four videos are mentioned in Table 2.

Table 2: Titles of TED Talks Videos

	Title	Speakers	Time (min)	Word numbers	Total
Native speakers	What 60 Schools can Tell Us About Teaching 21st Century Skills	Grant Lichtman	15:29	3190	6323
	How to Escape Education's Death Valley	Ken Robinson	19:11	3133	
Non-native speakers	Becoming a Better Teacher	Mariappan	19:54	3387	6135
	How School can Stop Teaching and Start Learning	AnjuMusafir	14:26	2748	

In order to answer the three research questions, the researcher carried out two steps as follows. In the first step, the researcher collected four TED talks videos (native and non-native speakers) from the website (www.Ted.com) with the topics related to education. The researcher then copied these four videos' content and stored them as "plain text" files for the AntConc analysis. In the second step, the researcher analyzed the frequency word lists according to four videos by listing the top 10, 20, and top 100 frequency word lists of the TED talks corpus of native and non-native speakers, separately.

3.2 Data Analysis

After collecting four TED talks videos (two from NS and two from NNS), the researcher put all the text files into *AntConc* software and used the word list tool to check the frequency word lists used by native speakers and non-native speakers, separately. The researcher then made a list of the top 10 and 20 frequency words of TED talks from native speaker videos and non-native speakers. In this case, the researcher expanded the ranking of word lists to 100 to see more clearly in the use of language between NS and NNS. Finally, the researcher categorized which words appear more than others and explored the differences in using language between NS and NNS. Figure 1 below summarizes the steps of data analysis.

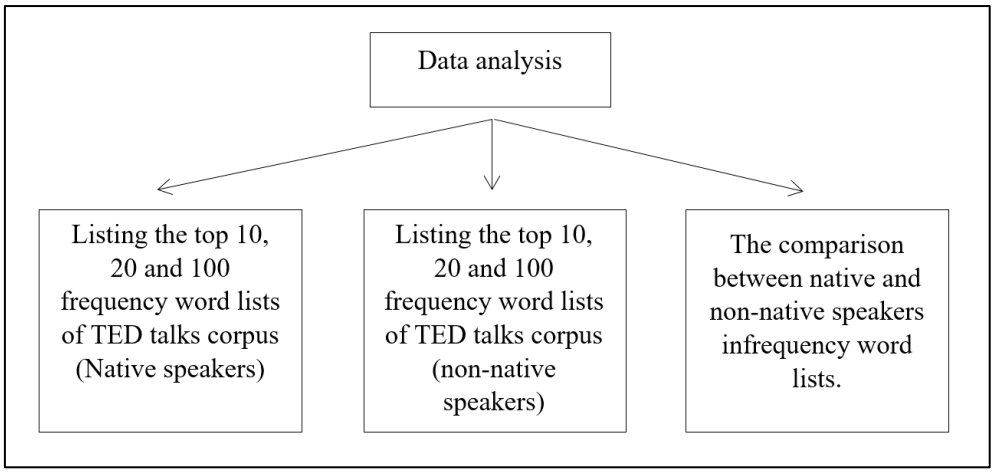

Figure 1: Data Analysis Steps

4. Results and Discussion

This part presents the word frequency lists analyzed by *AntConc* that the researcher collected from the TED talks videos. The findings are organized according to the three research questions. The first research question is about the frequency word lists in TED talks of native speakers in education topics. The content of the second research question is the same as the first question. However, the object is replaced by non-native speakers instead of native speakers. The last one is to explore the differences in the language used between people who use English as the first language and the ones who are successful users of English, but English is not their mother tongue.

4.1 The Word Frequency Lists in TED Talks of Native Speakers in Education Topic

There were 6323 running words of two native speakers TED talks transcript corpus and the list of 20 high-frequency order is illustrated in Table 3. It can be seen that the highest word frequency is "the," which appears more than 200 times among 6323 words. This result supports the finding of Wang's (2012) and Chen's (2017) studies when they also found "the" appears in the high frequency. It indicates that "the" is used very popular in speaking with all kinds of topics. The second to tenth high-frequency words are "and, of, that, to, a, I, we, it, is." These words appear more than 100 times in native TED talks corpus. The others in rank 11 to 20 are "in, s, you, they, t, this, what, have, not."

Table 3: Top 20 High Frequency Words of TED talks of Native Speakers

Rank	Frequency	Word numbers	Rank	Frequency	Word numbers
1	239	The	11	100	In
2	231	And	12	98	S
3	157	Of	13	72	You
4	154	That	14	66	Are
5	150	To	15	57	They
6	131	A	16	53	T
7	125	I	17	47	This
8	115	We	18	47	What
9	108	It	19	44	Have
10	102	Is	20	41	Not

4.2 The Word Frequency Lists in TED Talks from Non-native Speakers in Education Topic

Moving to the high-frequency words of TED talks from non-native speakers, it is exciting to note that "the" appears in the first high frequency, which is the same as the native TED talks corpus. The second to tenth high frequency are "and, to, I, you, a, of, that, is, in." These words are similar to the native TED talks corpus; however, the order between words is different. For example, "of, that" appear as top three and four, respectively in the native TED talks corpus, while, seven and eight in the non-native TED talks corpus (See Table 4 below).

Table 4: Top 20 High Frequency Words of TED Talks from Non-native Speakers

Rank	Frequency	Word numbers	Rank	Frequency	Word numbers
1	219	The	11	71	So
2	171	And	12	68	It
3	161	To	13	64	S
4	157	I	14	62	T
5	149	You	15	60	We
6	141	A	16	52	They
7	108	Of	17	43	Was
8	100	That	18	42	Are
9	97	Is	19	41	Students
10	95	In	20	41	What

4.3 The Comparison of Language Use between Native and Non-native Speakers in TED Talks

The occurrence of 10 words in Table 5 between native and non-native speakers indicates that "the, and" appear on top one and two in the frequency of word lists. Thus, it is noted that both native and non-native speakers prefer to use functional words in their speaking, however, "the" seems to be used by NS more than NNS (231 times compared to 171 times, respectively). Followed by "the, and" are "of, that, I, we, in, you…" These words appear in a different order among two different speakers.

Table 5: Top 10 High Frequency Words of TED Talks between Native and Non-native Speakers

a. Native Speakers		
Rank	**Frequency**	**Word**
1	239	The
2	231	And
3	157	Of
4	154	That
5	150	To
6	131	A
7	125	I
8	115	We
9	108	It
10	102	Is
b. Non-native Speakers		
Rank	**Frequency**	**Word**
1	219	The
2	171	And
3	161	To
4	157	I
5	149	You
6	141	A
7	108	Of
8	100	That
9	97	Is
10	95	In

"The" appears in the highest on the occurrence of word lists from native and non-native speakers; however, the number of occurrence of the word "the" in NS is more than NNS. The concordance lines of both NS and NNS corpus illustrate that most of the article "the" is combined with nouns or used in the comparative clause. Similarly, the conjunction "and" appears on the top two of the frequency word lists in both NN and NNS corpus. Using conjunction "and" is to connect between sentences, clauses, phrases, or words. When the sentences joined together by conjunctions, it could become meaningful.

The words "of, that" appear on the list of top ten in the frequency word lists from NS and NNS; however, the order is different as they are on the top three and four in the NN corpus; the top seven and eight in the NNS corpus, respectively. Meanwhile, the

pronouns "you, I" are used more frequently in NNS rather than NS did. Besides, looking more closely, it is interesting to note that NNS used the second-person pronoun "you" and NS used the first-person pronoun "we" in the top 10. It indicates that both NNS and NS usually use the first and the second person in their speech to share their personal experience.

According to Chung (2007), functional words are included "*pronouns, prepositions, articles, conjunctions, and auxiliary verbs*" (p. 347). In contrast, content words consist of nouns, verbs, adjectives, and adverbs. Thus, looking at Table 5 above, it can be seen that 100% functional words are used at surprisingly high rates in the top 10 word lists in both native and non-native speakers. Therefore, the research tried to expand the rank of the word lists to top 100 to see in detail between content and functional words. The result revealed that in the NS corpus, 23% content words and 77% functional words; in the NNS corpus, content words account for 29% in the top 100 words while 71% functional words. These numbers illustrate that functional words occupied almost double than content words. This result supported Chen's finding that there are 68% functional words and 32% content words in the TED talks corpus in business topics. Thus, we can include that people usually use functional words more than content words in all kinds of topics and different speakers.

In conclusion, functional words are used more frequently than content words. Nonetheless, the order of words is quite different among NS and NNS. The occurrence of more functional words on the top 10 or top 20, top 100 does not mean content words are not necessary. Because when the researcher looked at the concordance lines, it demonstrated that most functional words are combined with content words such as "we go…, we know… the American system… the arts…" Moreover, the other evidence from Wang's (2012) study when he concluded that the article "the" conveys no meaning when it stands alone; however, when it is combined with another word like "*look,"* it becomes a phrase that is very significant for connecting ideas "*look at the.*" Therefore, although the functional words always appear in the word lists' high frequency, we cannot deny content words' contributions. They play a pivotal role in making a full meaning sentence.

5. Conclusion

This part is composed of three sections. The first section is the purpose and significant finding of the current study. Next, the implications are presented. Finally, this study's limitations and suggestions for the next future studies are included in the third section.

The current study investigated the word frequency lists between NS and NNS from the TED talks source. The researcher collected four transcripts of education topics in TED talks videos. To avoid influencing the result during the process of calculating the frequency of word lists, the researcher tried to find videos that have almost the same length of time and the word numbers. However, the word numbers could not be precisely similar. *AntConc* is the main software used to generate the word frequency. The researcher found that "the" "and" are on the top one and two in NS and NNS corpus,

respectively. Although the functional words have little meaning, compared to content words, the figure of the top 100 frequency words in both NS and NNS corpus showed that the occurrence of functional words are almost double than content words.

Nevertheless, it is too hasty to include that content words are not valuable. It was apparent when the researcher looked at the concordance lines and saw that all functional words are connected with contents words to create a phrase or sentence. Thus, content words and functional words are inseparable, and they need to go hand in hand. Besides, in content words from both NS and NNS corpus, the researcher discovered that the words "teachers, students, education, teaching, learning" appear frequently in the top 100.

As for pedagogical implications, EFL teachers and students might find TED talks helpful in teaching and learning. For example, teachers might use TED talks as a real material in teaching *Public Speaking* course. Then, students can use TED talks as a source to practice listening and speaking skills. As for the connecting words, students not only find in the dictionary, but they also can find in TED talks corpus. Thus, it can be recognized that TED talks are considered as a useful source to look up English words. Moreover, based on this study's results, students need to learn how to use function words effectively because they are used in communication very often.

Finally, due to the limited number of TED talks corpus, time constraints, research design, and research questions, some limitations are inevitable when the study was conducted. For instance, in the first place, only four TED talks videos were collected to analyze, so the small number of videos in this current study might limit the result. Therefore, it is recommended that future research can increase the number of TED talks videos. Moreover, the research questions focus on the frequency of word lists only, so it is highly recommended that further research might expand it to more different aspects. Hence, it is hoped that future research will avoid these limitations.

Conflict of Interest Statement

The authors declare no conflicts of interests.

About the Author

Hieu Manh Do is a Lecturer in the Department of English Language at Hong Bang International University, Vietnam. He got MA TEFL at National Chung Cheng University, Taiwan in 2019. His main interests include English for Specific Purposes (ESP), English for Academic Purposes (EAP), Corpus Linguistics, Discourse Analysis, and English Language Skills.

References

Biber, D. and Jamieson, J. (1998). Final Report: Pilot study to test the influence of linguistic variables on listening and reading test performance. Princeton, NJ: Educational Testing.

Chen, I. C. (2017). *A Corpus-based study of the business-related TED talks from 2006 to 2016*. Southern Taiwan University of Science and Technology, Taiwan, Republic of China.

Chung, C., & Pennebaker, J. W. (2007). *The psychological functions of function words*. Social communication, 1, 343-359.

Corpus. Languaging Diversity: Identities, Genres, Discourses, 2015, 279-296.

Corpus-Driven Approach. International Journal of English Linguistics, 2(4), 33.

D'avanzo, Stefania. (2015). *Speaker identity vs. speaker diversity: The case of TED talks*

Flowerdew, L. (2011). *Corpora and language education*. Springer.

Hunston, S. (2002) Corpora in Applied Linguistics. Cambridge: Cambridge University Press.

Loan, C. V. (1990). How to have a good short talk. Retrieved April 15, 2012, from http://www.cs.cornell.edu/cv/ShortTalk.htm

O'keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Reppen, R. (2009). English language teaching and corpus linguistics: lessons from the American National Corpus. In P. Baker (Ed.), *Contemporary Approaches to Corpus Linguistics* (pp. 206-215). London: Continuum Press.

Sinclair, J. McH. (1991) Corpus, Concordance, Collocation. Oxford: Oxford University

Stubbs, M. (1996). Text and Corpus Analysis. Oxford: Blackwell.

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