

# An Exploration of Vocabulary Knowledge in English Short Talks

## *A Corpus-Driven Approach*

Yu-Chia Wang<sup>1</sup>

<sup>1</sup> Department of English, National Taiwan Normal University, Taiwan

Correspondence: Yu-Chia Wang, Department of English, National Taiwan Normal University, 162, Heping East Road Section 1, Taipei, Taiwan. E-mail: ycw233@gmail.com

Received: May 8, 2012

Accepted: May 28, 2012

Online Published: June 20, 2012

doi:10.5539/ijel.v2n4p33

URL: <http://dx.doi.org/10.5539/ijel.v2n4p33>

### Abstract

Adopting a corpus-driven approach, the study aimed to explore the vocabulary knowledge in English short talks including word patterns, features, and usages that are most likely to be encountered by language users in the real context. A specific corpus *TED* was conducted through a collection of English talks that are less than 20 minutes from the website *TED Talks*. In addition, the existed corpus *BASE* (British Academic Spoken English) was included in the study as a sample of talks longer than 20 minutes. Applying three corpus tools, *AntConc* (Anthony, 2003), *RANGE* (Nation & Heartkey, 2002), and *KfNgram* (Fletcher, 2007), the researcher was able to compile frequency-ordered word lists, concordance lines, vocabulary coverage, and lists of lexical bundles. The results showed that although the most frequently-used words in *TED* corpus and *BASE* corpus were similar grammatical items, the order was quite different. Moreover, the chi-square test showed a significant difference among four pronouns *I*, *You*, *We*, *They* between the two corpora and also in different parts of the *TED* corpus. Finally, the results of concordance lines and lexical bundles presented the “typical” and “frequent” word usages in the beginning, middle, and ending part of English short talks. It is suggested that teachers can build their own corpus to meet specific teaching purposes or learner’s needs, and to generate the corpus results into classroom materials while teaching English short talks.

**Keywords:** corpus-driven approach, English short talks, vocabulary

### 1. Introduction

For the past thirty years, corpus linguistics has been practiced a lot in the field of second language acquisition, providing teachers and researchers another way of choosing the more “authentic” and “communicative” materials in teaching and making the learning of language more fun and interesting. Under this framework, we are able to investigate the “language” native speakers practice in both written and spoken forms. According to McCarthy (2001, p. 125), corpus linguistic not only provides a way to engage learners in the “real” language, but will “impinge upon our long-held notions of education, roles of teachers, the cultural context of the delivery of educational services and the mediation of theory and technique.”

In a university context in Taiwan, asking students to make presentations is a popular method to evaluate their performance in a language classroom. The purpose of asking university students to give a short talk to the class, whether it is individual work or group work, is often to develop their language abilities, in particular their speaking skills and to prepare them for future careers. However, students are not born with the skills to give a public talk or speech, not mentioning they have to present it in a foreign language. Moreover, in order to give an appropriate public talk in a target language, not only does one need the knowledge of linguistic rules, he or she is also required to adopt the proper “register”, which is defined as the usage of a language for a specific purpose or for a particular setting. Therefore, before asking students to give a short talk or to present ideas in a target language such as English, teachers need to first point out the correct form of register or grammatical and lexical features that are unique for an English short talk. Only by investigating the “authentic” language used by native speakers or “frequent” users of that target language beforehand can teachers provide their students sufficient information about the kind of skills they need for demonstrating the challenging task.

In order to provide adequate data for future teachers and learners, a computer-mediated corpus-driven approach is encouraged in eliciting useful details. For example, Biber et al. (1999) applied different corpus analysis tools in their study and was able to describe the linguistic features of four registers in English which were also the most frequently encountered ones by native speakers. Moreover, their findings provided language teachers and students valuable information to design a syllabus, to write teaching materials, and to use as a reference to study English conversation, newspapers, fiction, and academic prose. Therefore, by collecting and investigating a large amount of “real” English talks by professional presenters, the present study aims to build a specific “corpus” that will help us to interpret the insightful linguistic properties of English short talks. Hopefully, by analyzing the corpora with existed corpus tools, the results will also provide English teachers and students a better understanding of how the language is being used in giving English short talks.

## 2. Literature Review

### 2.1 Corpus-based vs. Corpus-driven Study

For many years, researchers have conducted corpus studies to examine the real quality of the use of the target language and some of them even compared the result with textbooks, the one we used to count on with 100% confidence. For example, Holmes (1988) compared modal verbs in ESL textbooks with corpus data and found a gap between the textbook information and how the language is really being used. Boxer and Pickering (1995) compared the dialogues in textbooks with spontaneous real speech and suggested that class materials should not rely only on native speaker’s intuition which might be faulty sometimes. All of these empirical studies proved the important role of corpus linguistics in assisting second language teaching and learning.

However, before we start to work on a corpus study, two general approaches should be distinguished- “corpus-based” and “corpus-driven” studies. According to Tognini-Bonelli (2001), a “corpus-based” study assumes a pre-existing theory of the language usage. The purpose of a corpus-based study is always to test whether the pre-defined linguistic rule is valid by analyzing the corpora data. Therefore, the outcome of corpus-based research is usually deductive, standardized, and simplified. On the other hand, a “corpus-driven” study can be more inductive. In a corpus-driven study, researchers usually allow linguistic features to emerge from the naturally occurring context, “exploiting the potential of a corpus to identify linguistic categories and units that have not been previously recognized” (Biber, 2009, p. 278).

In the present study, since the goal was not to test or challenge any existing linguistic theories, nor did the researcher assume any pre-established language rules for English short talks, an inductive corpus-driven approach was adopted to uncover the unknown characteristics of English short talks.

### 2.2 A Definition of English Short Talks

For a long time, researchers assumed a dichotomous relationship between written and spoken language and conducted studies from different perspectives; for example, the comparison of lexical features, sentence structures, sentence length and word syllables in spoken and written language (Drieman, 1962; Gibson et al., 1966; Kroll, 1977). However, according to Cleland and Pickering (2006, p. 185) “the relationship between speaking and writing has sometimes been taken for granted”. Moreover, “the dangers of too easy an acceptance of such a dichotomy are worth repeating here, even if they are obvious” (Newman, 2010, p. 83). For instance, the lexical patterns of a presidential speech may be more similar to a written work rather than a spoken one.

Similarly, we can never assume that preparing for a speech for a group of scientists will be the same as preparing for the speech for a class of first-year college science majors or that giving a 20 minute presentation will be the same as giving a 50 minute one. Therefore, in order to explore the knowledge of vocabulary in English short talks, we need to first define what an English short talk is. Adopting from Loan’s (1990) definition, a short talk is a less than 20 minute presentation that “requires a more sustained level of clarity if it is to be successful”; moreover, “like any presentation, a short talk should have a beginning, a middle, and an end”. In the present study, the transcripts of 30 presentations that are less than 20 minutes in TED Talks were selected and used to build the TED corpus.

### 2.3 Previous Corpus Studies & L2 Learning

For decades, corpus linguistics has been seen as a strongly empirical methodology to help us reveal language changes, language development, and language in use. For example, Hughes and McCarthy (1998) looked at the use of past perfect verb forms by native speakers in CANCODE, and found that the use of past perfect forms had a more complex function in spoken discourse than it was listed in textbooks. Cacoullos and Walker (2009) used the variationist method to examine the various use of “will” and “going to”, and concluded that “the choice of form is not determined by invariant semantic readings such as proximity, certainty, willingness, or intention.

Rather, particular instances of each general construction occupy lexical, syntactic, and pragmatic niches” (p. 321). Moreover, Nesi and Basturkmen (2006) in their study found that four word lexical bundles (words must be used together such as *I don't think...would you mind....*) played a discourse signaling role in lectures that was crucial for language learners to be aware of. They argued:

When native speakers of English can be expected to have implicit knowledge of the function of bundles, non native speakers are much less likely to have this understanding because they have consciously learned the language, rather than acquired it, and the role of lexical bundles as discourse signals is yet to be acknowledged in most language teaching materials. (p. 300).

Moreover, the application of corpus study result has also been suggested in language teaching and learning. For instance, Johns (1986) recognized the values of concordancing tools in second language learning, especially for teaching English for Specific Purpose (ESP). He proposed that by working with a concordance, students were able to study the appropriate word usages in a way that was more effective than traditional class procedures. Reppen (2009) argued that with the concordancing program MonoConc, teachers could develop their own teaching materials and activities to help learners to identify the multiple meanings of the word *like*. Furthermore, he suggested teachers and researchers to examine texts across different registers because they are “created for different purposes under different conditions”, and “have different linguistic features associated them” (p. 209).

To conclude, by conducting corpus studies, corpus-based or corpus-driven, researchers were able to validate pre-existed theories on the grammatical and lexical functions of the language, or to relate certain linguistic features and usages of the target language, whether the purpose of the study aimed to look at individual words, to study different genres, the organization of culture and social purposes around language that is tied closely to the ideology and power (Bhatia, 1993; Swales, 1990), or to consider the different registers under different situations or contexts. Therefore, in order to make contribution to the teaching and learning of English short talks to non-native speaking learners, the present study adopted a corpus-driven approach that aimed to reveal the patterns, features, and usages in the specific corpora with a variety of corpus analysis tools, hoping to “provide a rich resource for teachers preparing students for a particular context of English use” and for specific teaching purposes (Reppen, 2009, p. 207).

### 3. Methodology

#### 3.1 Data Collection

The current study tended to collect one specific domain of data-English short talks from TED ([www.ted.com](http://www.ted.com)), a non-profit organization “began as a simple attempt to share what happens at TED with the world, under the moniker ideas worth spreading”, and one existed corpus data-The British Academic Spoken English (BASE). The BASE corpus was downloaded from the website of the university of Warwick and Reading; the TED corpus was a collection of thirty transcripts from TED’s web site. Applying different corpus tools, the study aimed to uncover the features or patterns of English vocabulary usages in English short talks, and to make suggestions for teaching and learning how to give English short talks.

##### 3.1.1 BASE

The BASE corpus was taken into to see if there was a difference between the words used in English short talks (talks that are less than 20 minutes) and talks that are more than 20 minutes. In the BASE corpus, speeches related to physical science and social science were chosen in order to be comparable to the themes in the TED corpus. Forty lectures and ten seminars of each theme were selected with an average length between 50 to 100 minutes (Table 1).

Table 1. Descriptions of BASE corpus

| Theme              | Physical Science | Social Science |
|--------------------|------------------|----------------|
| Number of Lectures | 40               | 40             |
| Number of Seminars | 10               | 10             |
| Total tokens       | 643,649          |                |

##### 3.1.2 TED Talks

Thirty TED talks were chosen from three themes-science & technology, global issues, and business, with an average length of sixteen to eighteen minutes (Table 2).

Table 2. Descriptions of TED corpus

| Theme           | Science & Technology  | Global Issues      | Business         |
|-----------------|-----------------------|--------------------|------------------|
| Number of Talks | 10                    | 10                 | 10               |
| Average Lengths | 16 minutes 44 seconds | 16 minutes 44 sec. | 17minues 20 sec. |
| Total Tokens    | 80,885                |                    |                  |

### 3.2 Data Analysis

#### 3.2.1 Word-listing & Concordancing

Word-listing is a basic technique to present words of a text in a systematical way. Based on the purpose of the study, words can be transformed in alphabetical orders, frequency orders, reverse alphabetical orders, or word-length orders (Scott & Tribble, 2006). A word list is also essential for word classification. In the present study, a frequency word-listing technique was adopted. As a result, word appeared most frequently in the text appeared on the top in the word list, followed by the less frequent words in the text. However, previous researchers argued that with this method, usually the most frequent words found were functional words such as *the, of, and, to, a, in*, etc., and they were often not very informative. On the other hand, Scott and Tribble (2006) have suggested that while comparing two different texts, even though we could only identify functional words, if two word lists were not with identical order, “it would be worth investigating whether the slight differences are (1) typical of the most frequent words or not, (2) illustrative of the different text types, or (3) representations of language changes” (Scott & Tribble, p. 17). In other words, even if a frequency ordered word list may contain all grammatical words that are irrelevant to the text content, they may still carry important message that is worth exploring.

In the current study, AntConc (Anthony, 2003), a free concordancing software was used to compile the frequency ordered word lists and concordance lines. Concordancing, “a process of using software to search for all the occurrences of one word (or phrase) in a corpus” (O’Keeffe & Farr, 2003, p. 393) is efficient to find the “grammatical and collocation patterns that emerge for the word” (p. 394). Meanwhile, it was also suggested that using concordance lines in teaching and learning a second language can help raise students’ awareness of grammatical and lexical patterns and develop their problem-solving skills especially in classroom activities (Fox, 1998; Johns, 1997; Stevens, 1995).

#### 3.2.2 Vocabulary Coverage

While examining the difficulty of a text or to define how much a learner can understand the text, researchers often paid attention on vocabulary coverage, including word levels and percent coverage. For example, Laufer (1992) found that approximately 3,000 words were required for reading texts that are at the university level, while 5,000 words were claimed for academic success. Nation (2006) argued that 98 % coverage of vocabulary should be reached for the comprehension of television programs. Moreover, according to Webb and Rodgers (2009), the most frequent 4,000 word families covered 95 % vocabulary in general American programs, while the most frequent 8,000 word families covered 98% vocabulary in it. The result suggested that in order to understanding general American TV programs, learners may need a vocabulary size of 6,000 to 8,000 words families.

Adopting RANGE (Nation & Heartkey, 2002), a computer program that lists the coverage and level of word families in texts, the present study compared vocabulary coverage in both English short talks and academic speeches to see if there was a gap between the two types of spoken language.

#### 3.2.3 Lexical Bundles

Many studies have done in recent years to investigate lexical bundles (or n-grams and clusters), which are words that frequently appear together that may be recognized as common phrases or a special combinations of words (Biber, et al., 2004; Hoey, 2005; Carter & McCarthy, 2006). For example, Carter and McCarthy found specific usages of lexical bundles in spoken language that reflected “interpersonal meaning” such as *you know, I know what you mean, and I think* (as cited in Greaves & Warren, 2010, p. 216). Moreover, they identified frequent lexical bundles used to express vagueness in spoken language such as *kind of, something like that, and all the rest of it*. According to Greaces and Warren (2010), the finding of multi-word units in a corpus facilitated language teaching and learning because most texts were made of common words with common patterns in that target language. Besides, compared to multi-word units, words usually have no independent meaning when they appear individually in a text.

Therefore, the current study adopted KfNgram (Fletcher, 2007), a free software program that helps to generate lists of lexical bundles in texts, and aimed to find the most frequent combinations of words, the lexical bundles used in English short talks.

#### 4. Results & Discussion

##### 4.1 The Most Frequent Words in TED and BASE

In order to compare words that appear in TED and BASE, two word lists were compiled using AntConc tools. Table 3 and Table 4 showed the top ten most frequent words in the TED corpus and BASE. As previous studies suggested, the most frequent words in the list were function words such as *the*, *to*, *a*, *that*, *and*, etc. However, a slight difference could be identified between the two corpora according to the ranks. For example, although *of*, *to*, are both on the top ten lists, their ranks are contrastive. Other function words like *that*, *and* and *show* similar results.

Table 3. Frequency ordered word list of TED

| Rank | Word | Frequency |
|------|------|-----------|
| 1    | THE  | 3678      |
| 2    | TO   | 2235      |
| 3    | OF   | 2146      |
| 4    | A    | 1922      |
| 5    | THAT | 1721      |
| 6    | AND  | 1496      |
| 7    | IN   | 1386      |
| 8    | IS   | 1377      |
| 9    | YOU  | 1127      |
| 10   | IT   | 1088      |

Table 4. Frequency ordered word list of BASE

| Rank | Word | Frequency |
|------|------|-----------|
| 1    | THE  | 33691     |
| 2    | OF   | 17976     |
| 3    | TO   | 16802     |
| 4    | AND  | 16311     |
| 5    | THAT | 14670     |
| 6    | A    | 13767     |
| 7    | YOU  | 13002     |
| 8    | ER   | 12924     |
| 9    | IN   | 11226     |
| 10   | IT   | 10843     |

Next, using RANGE program, vocabulary coverage was revealed. First of all, in the TED talks, about 91% (84.86% + 6.05%) of the words used by the presenters was sorted in base list one and two (the first 2,000 most frequent words of English), and 93 % ((84.86% + 6.05% + 2.04%) of the words were sorted in the base list one, two, and three, meaning that with a 2000 word level vocabulary size, one could understand 91% of the words in TED talks; moreover, a person with a 3000 word vocabulary size could recognize 93% of the words in TED talks (Table 5). However, according to the BASE corpus, a person with a 2000 word level vocabulary size only has a chance to understand about 89% (83.67 + 4.45%) of the lectures and seminars; while a person with a 3000 word level vocabulary size can understand 90.65% (83.67 + 4.45% + 1.53) of academic speech (Table 5).

Table 5. Vocabulary coverage of TED and BASE

|                         | <b>TED<br/>(Tokens/%)</b> | <b>BASE<br/>(Tokens/%)</b> |
|-------------------------|---------------------------|----------------------------|
| <b>1000 word level</b>  | 68,382/ 84.86%            | 531,329/ 83.67%            |
| <b>2000 word level</b>  | 4,878/ 6.05%              | 34,581/ 5.45%              |
| <b>3000 word level</b>  | 1,644/ 2.04%              | 9,747/ 1.53%               |
| <b>Not in the lists</b> | 5677/ 7.05%               | 59340/ 9.34%               |

However, in order to have a closer examination into the relationship between TED and BASE, four high frequency pronouns: *I*, *you*, *we*, and *they* were chosen for comparison using chi-square test. The purpose of the chi-square test was to estimate whether the frequencies of the four pronouns differ significantly from each other in both corpora. Table 6 presented the frequencies of the four pronouns in both corpora.

Table 6. Pronoun frequencies for two different corpora

|             | <b>TED</b> | <b>BASE</b> | <b>Total</b> |
|-------------|------------|-------------|--------------|
| <b>I</b>    | 932        | 7658        | 8590         |
| <b>YOU</b>  | 197        | 1127        | 1324         |
| <b>WE</b>   | 13002      | 44          | 13046        |
| <b>THEY</b> | 434        | 4603        | 5037         |

Moreover, Figure 1 showed that the four types of pronouns in the TED and the BASE were significantly different with a less than .01 alpha level, meaning that there was only a 1% probability that the result occurred by chance alone.

| <b>Chi-Square Tests</b>      |                        |           |                              |
|------------------------------|------------------------|-----------|------------------------------|
|                              | <b>Value</b>           | <b>df</b> | <b>Asymp. Sig. (2-sided)</b> |
| Pearson Chi-Square           | 22231.458 <sup>a</sup> | 3         | .000                         |
| Likelihood Ratio             | 28207.303              | 3         | .000                         |
| Linear-by-Linear Association | 3385.204               | 1         | .000                         |
| N of Valid Cases             | 27997                  |           |                              |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 635.21.

Figure 1. Chi-square test of pronoun frequencies for two corpora

In sum, the finding suggested that although both corpora contained similar grammatical words from their frequency ordered word lists, minor distinctions could be identified in their orders. In addition, comparing to an academic speech, the vocabulary level and the percentage of word coverage in English short talks seemed to be lower and lesser. Although the possible explanation could be that speeches in the BASE corpus were targeting at mostly university level students in the specific field, where short talks in the TED corpus were aiming for more general audiences, the result seemed to reject previous argument: "If you can give good a short talk then you can probably give a good 50-minute presentation because the additional time permits certain flexibility" (Loan, 1990).

#### 4.2 The Most Frequent Words Appear in the Beginning, Middle, and Ending Part of TED

As shown in Table 4, the most frequent words in TED Talks were grammatical words such as articles, pronouns, or prepositions. In order to explore more about how these words were used in English short talks, three different word lists were compiled based on the parts in which the words appeared. The purpose was to see how these high frequency words differ in the *beginning*, *middle* and *ending* parts of the talks. Therefore, the *beginning* part was a collection of all the words appeared in the first paragraph in TED talks. The *ending* was a collection of all the last paragraphs in the talks, and the *middle* was a collection of all the paragraphs except for the first and the last ones in all talks. Table 7 showed the frequencies and percentages of the four pronouns *I*, *You*, *We*, *They*, in three parts.

Table 7.

|             | <i>Beginning</i>         | <i>Middle</i>            | <i>Ending</i>            |
|-------------|--------------------------|--------------------------|--------------------------|
|             | Frequencies/ Percentages | Frequencies/ Percentages | Frequencies/ Percentages |
| <b>I</b>    | 53/2.45%                 | 895/1.1%                 | 26/1.16%                 |
| <b>YOU</b>  | 23/1.06%                 | 1061/1.4%                | 57/2.55%                 |
| <b>WE</b>   | 26/1.2%                  | 1022/1.34%               | 56/2.5%                  |
| <b>THEY</b> | 6/0.27                   | 405/0.53%                | 8/0.36%                  |

As shown in Table 7, the pronoun *I* appeared in the *beginning* part of the talks with a percentage of 2.45, which is higher than when it appeared in both the *middle* and *ending* part of the talks. Next, *You* and *We* appeared in the *ending* with a percentage of 2.55 and 2.5 that was also higher than when they appeared in both *beginning* and *middle* parts of the talks. Then, *they* appeared in the *middle* with a percentage of 0.53 that was higher than when it appeared in both *beginning* and *ending* parts. Finally, the chi-square test result showed that that the four types of pronouns appeared significantly ( $p < .01$ ) different in the *beginning*, *middle*, and *ending* part of the short talks (Figure 2).

| Chi-Square Tests             |                     |    |                       |
|------------------------------|---------------------|----|-----------------------|
|                              | Value               | df | Asymp. Sig. (2-sided) |
| Pearson Chi-Square           | 43.766 <sup>a</sup> | 6  | .000                  |
| Likelihood Ratio             | 42.484              | 6  | .000                  |
| Linear-by-Linear Association | 10.471              | 1  | .001                  |
| N of Valid Cases             | 3638                |    |                       |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.44.

Figure 2. Chi-square test of pronoun frequencies in the *beginning*, *middle*, and *ending* parts

Overall, the result suggested that in English short talks, presenters used the pronoun *I* more often in the beginning of the talks, while *you* and *we* were mostly used in the *ending* and *they* was largely used in the *middle* part of the talks. In addition, in order to know how these pronouns were actually used in short talks, concordance lines were compiled and interpreted. For example, Figure 3 showed that pronoun *I* appeared mostly in the *beginning* part of the talk to carry out personal experiences such as *when I was five years old*; *I always wondered*; and *I particularly remember*, or to introduce the purpose of the talks like *I'd like to talk* and *I'd like to discuss*. Moreover, as shown in Figure 4, pronoun *you* mostly appeared in the *ending* to motivate the audiences, for instance, *if you ever*; *I would leave you with*; *I hope you will*, or to simply thank the audiences as in *Thank you very much*. Next, Figure 5 revealed that pronoun *we* also appeared mostly in *ending*, and was used to urge for an action in the future like *we have to*; *if we can*; and *if we can't*. Finally, Figure 6 showed that pronoun *they* occurred mostly in the *middle* part while giving examples or elaborating ideas such as *they told*; *they showed up*; and *they begin to*.

```

at play here. When I was five years old I fell in love with ai
as five years old I fell in love with airplanes. Now I'm talk
th airplanes. Now I'm talking about the '30s. In the '30s an
t time. Of course I always wondered what would happen if he'd
he airplane first. I've always been interested in relationship
ehavior. And what I'd like to discuss today is the need to ov
eft my brain. And I started playing with it more like a puzzl
original thought I had, this must be ethanol. So I went out
st be ethanol. So I went out and researched ethanol. And four
six months later I figured out it must be hydrogen, until so
advertising man, I actually speak at TED Evil, which is TED'
ars in Burma. And I particularly remember a really good spec
ens smoking again. I'd like to talk a little bit this morning
the first project I was ever hired to do. Something like 25 y
it easier to use. I thought, at the time, I did a pretty good
ght, at the time, I did a pretty good job. Unfortunately, not
innovation? Now, I want to tell you a quick story. We'll go
fact, the date -- I'm curious to know if any of you know what
-----

```

Figure 3. Concordance lines containing *I* in the *beginning* part of the TED corpus

Note: Generated with AntConc Tools (Anthony, 2003).

I have. Thanks very much. Thank you very much. So one of these ta  
in them, and she's right. So if you ever, ever get an opportuni  
turn out the lights. I promise, you'll love it. Thank you. NASA  
promise, you'll love it. Thank you. NASA has this phrase that t  
novation was done without risk. You have to be willing to take  
hat's the thought I would leave you with, is that in whatever y  
e you with, is that in whatever you're doing, failure is an opt  
nk you. So, I hope I've convinced you of this -- of the impact on  
. Thanks for your attention. So, you know, that's my last messag  
know, that's my last message to you. How do we set a going-to-t  
ing to move nowhere. So, I hope you'll turn this conference int  
ate your attention today. Thank you. Anthony Atala: See, at the  
make our patients better. Thank you for your attention. It's a l  
ite the textbook about Mars. If you're interested in more infor

Figure 4. Concordance lines containing *you* in the *ending* part of the TED corpus

Note: Generated with AntConc Tools (Anthony, 2003).

merica. Thank you very much. If we can keep innovating on our spa  
h rules for changing rules, so we don't get stuck with bad rules  
get stuck with bad rules, then we can keep moving progress forwa  
te of this session, which is, "We are perishing for want of won  
t products. To do that I think we have to take a more expansive  
these thoughts. First of all: we all form tribes, all of us. Yo  
nged the world. If you do what we've talked about, you listen fo  
l five culture stages. Because we've got people in all five, arc  
tion of our human dignity that we would all say is horrific it's  
is horrific it's slavery. And we've got to say, what good is al  
lectual power in this room, if we can't use it to bring slavery  
an end. And you know what? If we can't do that, if we can't use  
what? If we can't do that, if we can't use our intellectual pot  
here is one last question, are we truly free? Okay, thank you s  
ch." They said, "Let's march." We should be marching towards a

Figure 5. Concordance lines containing *we* in the *ending* part of TED corpus

Note: Generated with AntConc Tools (Anthony, 2003).

d they made it their own, and they told people. And some of thos  
them showed up for him? Zero. They showed up for themselves. It  
up for themselves. It's what they believed about America that  
e middle of August. It's what they believed, and it wasn't about  
comprehensive 12-point plans. They're not inspiring anybody. Bec  
who lead inspire us. Whether they're individuals or organizatio  
And I went in to art because they appreciated drawing. Studied  
dn't choose airplanes because they had gotten sort of unromantic  
ng. And I felt, the one thing they don't need, is a chair that  
hanistically possible so that they didn't have to fuss with it.  
le with a lot of bulk up top. They begin to fall off the end of  
le don't adjust their chairs. They will sit in them forever. I l

Figure 6. Concordance lines containing *they* in the *middle* part of TED corpus

Note: Generated with AntConc Tools (Anthony, 2003).

#### 4.3 The Most Frequent Lexical Bundles in TED

To answer this question, KfNgram program was adopted in sorting out the most frequent three, four, and five-word lexical bundles in TED Talks. Table 8 showed the results of the most frequent three-word lexical bundles in TED corpus in three types. The first type included common phrases such as *a lot of*; *in order to*; *a little bit*; and *a couple of*. The second type contained words that were used to "interact" with the audiences such

as *I'd like to*; *we have to*; and *we need to*. The last type covered words to illustrate the speakers' slides, charts, or any other data such as *## percent of*; *you look at*; *here is the*; and *out of the*.

Table 8. Three-word lexical bundles in TED talks

| Rank | Lexical Bundles   | Rank | Lexical Bundles | Rank | Lexical Bundles |
|------|-------------------|------|-----------------|------|-----------------|
| 1    | A LOT OF          | 14   | THE FACT THAT   | 27   | HERE IS THE     |
| 2    | ONE OF THE        | 15   | YOU HAVE TO     | 28   | THE SAME TIME   |
| 3    | THIS IS A         | 16   | LOOK AT THE     | 29   | I DON'T KNOW    |
| 4    | THERE IS A        | 17   | PART OF THE     | 30   | ONE OF THOSE    |
| 5    | A LITTLE BIT      | 18   | SOME OF THE     | 31   | ALL OF THE      |
| 6    | ## PERCENT OF     | 19   | THE KIND OF     | 32   | BACK TO THE     |
| 7    | YOU CAN SEE       | 20   | THE REST OF     | 33   | IN ORDER TO     |
| 8    | WE HAVE TO        | 21   | ## YEAS AGO     | 34   | OUT OF THE      |
| 9    | WE NEED TO        | 22   | THE END OF      | 35   | THE POWER OF    |
| 10   | THE UNITED STATES | 23   | THE FIRST TIME  | 36   | THE WAY WE      |
| 11   | AROUND THE WORLD  | 24   | WHAT YOU DO     | 37   | YOU LOOKAT      |
| 12   | IN THE WORLD      | 25   | BE ABLE TO      | 38   | I'D LIKE TO     |
| 13   | A COUPLE OF       | 26   | DON'T WANT TO   | 39   | THE LAST ##     |

## 5. Conclusion

The current study showed a way to apply corpus-driven approach to explore English short talks. By building a specific corpus-TED, and comparing the result with the existed corpus-BASE, the study aimed to provide implications for English learners and teachers for a specific purpose, for example, learning about English short talks.

First of all, the comparison of word frequency lists and vocabulary coverage showed that there was a difference between words in the corpus of short talks and the corpus of academic speeches. Although in both corpora, the most frequent words were the same functional words, the orders were varied. Moreover, when further examining the four pronouns *I*, *you*, *we*, and *they* between the two corpora, a significant difference was found. Therefore, although short talks and academic speeches are labeled as spoken language, they seem to be lexically different. The result suggested a necessity to identify the differences between preparing for a 20 minute short talk and performing an over 50 minute speech.

In addition, evidences showed that the same words could function very differently even within the same corpus. For example, the chi-square test showed that the four pronouns *I*, *you*, *we*, and *they* were significantly different in the beginning, the middle, and the ending part of the short talks. Moreover, the concordance lines demonstrated examples of how the pronouns functioned differently in three parts of the talks while carrying out the topics by sharing personal experiences, elaborating the main ideas by talking about relevant stories, or making a powerful conclusion while inviting the audiences to "change" something in the future. The result suggested the necessity of identifying the different word usages in the beginning, middle, and ending parts of English short talks.

Finally, the KfNgram program was conducted to elicit re-occurring lexical bundles in short talks, and provided authentic word data for the learning of English short talks. Moreover, while investigating the multi-unit words in the corpus, learners and teachers were able to construct the "meaning" of the language used in English short talks. For example, the article *THE* and the preposition *OF* were both ranked top ten high frequency words in TED talks. When we look at them as individual words, they mean nothing but the so-called grammatical words; however, when they are combined with another word like *LOOK*, *PART*, *KIND*, or *REST*, they become phrases that is essential for connecting ideas in a talk such as *LOOK AT THE*, *PART OF THE*, *THE KIND OF*, and, *THE REST OF*. According to Hoey (2005, p. 8), "Our knowledge of a word includes the fact that it co-occurs with certain other words in certain kind of context." Therefore, the finding of the lexical bundles provided teachers and learners a deeper and more comprehensive way of learning how the specific words could be combined and functioned in a particular context.

## 6. Limitations

In the current study, while the lengths, themes, and dates in the TED talks were carefully selected and controlled to be comparable to another existed corpus, the BASE, the definition of English short talks was also limited since only data from the TED website were chosen. Therefore, it is not appropriate to generate the result to all kinds of English short talks, for example, the kind of talk given by a boss who tries to motivate his employees in a regular meeting or the kind of “instructional” talk a teacher uses to lecture her elementary school students. Here, according to the definition provided by TED, short talks involve a person’s intention to share “ideas worth spreading” in the world.

Moreover, although the current study was able to identify the unique features and usages of vocabulary in English short talks, it takes more effort to transform the findings into useful information for teaching and learning purposes. Only by understanding learners’ needs and the goals of the classes can we make research findings more effective to our students and can the learning become more meaningful. Hopefully, the result of the study will provide pedagogical implications to not only language teachers and learners, but also future curriculum designers, and policy makers.

## References

- Bhatia, V. K. (1993). *Analysing Genre: Language Use in Professional Settings*. London: Longman.
- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371-405. <http://dx.doi.org/10.1093/applin/25.3.371>
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson Education.
- Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3), 275-311. <http://dx.doi.org/10.1075/ijcl.14.3.08bib>
- Cacoullos, R. T., & Walker, J. A. (2009). The present of the English future: Grammatical variation and collocations in discourse. *Language*, 85(2), 321-354. <http://dx.doi.org/10.1353/lan.0.0110>
- Carter, R. A., & McCarthy, M. J. (2006). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.
- Cleland, A. A., & Pickering, M. J. (2006). Do writing and speaking employ the same syntactic representations? *Journal of Memory and Language*, 54, 185-198. <http://dx.doi.org/10.1016/j.jml.2005.10.003>
- Deroey, K. (2009). Corpus-informed EAP syllabus design: a study of lecture functions. Paper presented at BAAHE (Belgian Association of Anglicists in Higher Education), Namur, Belgium.
- Drieman, G. H. J. (1962). Difference between written and spoken languages: an exploratory study. *Acta Psychol*, 20, 36-57. [http://dx.doi.org/10.1016/0001-6918\(62\)90006-9](http://dx.doi.org/10.1016/0001-6918(62)90006-9)
- Farr, F., & O’Keeffe, A. (2003). Using language corpora in initial teacher education: pedagogic issues and practical applications. *TESOL Quarterly*, 37(3), 389-418. <http://dx.doi.org/10.2307/3588397>
- Flowerdew, L. (2003). A Combined Corpus and Systemic-Functional Analysis of the Problem-Solution Pattern in a Student and Professional Corpus of Technical Writing. *TESOL Quarterly*, 37(3), 489-511. <http://dx.doi.org/10.2307/3588401>
- Fox, G. (1998). Using corpus data in the classroom. In B. Tomlinson (Ed.), *Material development in language teaching* (pp. 25-43). Cambridge: Cambridge University Press.
- Gibson, J. W., Gruner, C. R., Kibler, R. J., & Kelly, F. J. (1966). A quantitative examination of differences and similarities in written and spoken messages. *Speech Monographs*, 33, 444-451. <http://dx.doi.org/10.1080/03637756609375510>
- Greaves, C., & Warren, M. (2010). What can a corpus tell us about multi-word units? In A. O’Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 212-226). Oxford: Routledge.
- Hoey, M. (2005). *Lexical Priming: A new theory of words and language*. London: Routledge.
- Holmes, J. (1988). Doubt and certainty in ESL textbooks. *Applied Linguistics*, 9, 21-44. <http://dx.doi.org/10.1093/applin/9.1.21>
- Johns, T. (1986). Micro-concord: A language learner’s research tool. *System*, 14(2), 151-162. [http://dx.doi.org/10.1016/0346-251X\(86\)90004-7](http://dx.doi.org/10.1016/0346-251X(86)90004-7)

- Johns, T. (1997). Context: The background, development and trialling of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100-115). London: Longman.
- Kroll, B. (1977). Combining ideas in written and spoken English. In E. O. Keenan, & Bennett, T. L. (Eds.), *Discourse across time and space* (pp. 69-108). Los Angeles: Dept. Linguist. Univ. South. Calif.
- Laufer, B. (1992). How much lexis is necessary for reading comprehension? In H. Bejoint, & P. Arnaud (Eds.), *Vocabulary and applied linguistics* (pp. 126-132). London: MacMillan.
- Loan, C. V. (1990). How to have a good short talk. Retrieved April 15, 2012, from <http://www.cs.cornell.edu/cv/ShortTalk.htm>
- McCarthy, M. J. (2001). *Issues in Applied Linguistics*. Cambridge: Cambridge University Press.
- Nesi, H. (2001). A corpus based analysis of academic lectures across disciplines. In Cotterill, J., & Ife, A. (Eds.), *Language across boundaries* (pp. 201-218). London: Continuum Press.
- Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signaling in academic lecturers. *International Journal of Corpus Linguistics*, 11(3), 283-304. <http://dx.doi.org/10.1075/ijcl.11.3.04nes>
- Newman, J. (2010). Balancing acts: Empirical pursuits in cognitive linguistics. In Dylan Glynn, & Kerstin Fischer (Eds.), *Quantitative Methods in Cognitive Semantics* (pp. 79-100). Berlin and New York: Mouton de Gruyter. <http://dx.doi.org/10.1515/9783110226423.79>
- O'Keeffe, A., & Farr, F. (2003). Using language corpora in initial teacher education: pedagogic issues and practical applications. *TESOL Quarterly*, 37(3), 389-418. <http://dx.doi.org/10.2307/3588397>
- 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.
- Scott, M., & Tribble, C. (2006). *Textual Patterns: keyword and corpus analysis in language education*. Amsterdam: Benjamins.
- Simpson, R., & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly*, 37(3), 419-441. <http://dx.doi.org/10.2307/3588398>
- Stevens, V. (1995). Concordancing with language learners: Why? When? What? *CAELL Journal*, 6, 2-10.
- Swales, J. (1990). *Genre analysis: English in academic and research settings*. New York: Cambridge University Press. [http://dx.doi.org/10.1016/0889-4906\(90\)90032-8](http://dx.doi.org/10.1016/0889-4906(90)90032-8)
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam/Philadelphia: John Benjamins.
- Webb, S. & Rodgers, M. P. H. (2009). Vocabulary demands of television programs. *Language Learning*, 59(2), 335-366. <http://dx.doi.org/10.1111/j.1467-9922.2009.00509.x>