THE TED WORD LIST: AN ANALYSIS OF TED TALKS TO BENEFIT ESL TEACHERS AND LEARNERS

By
JASON DANIEL WOLFE

B.Sc., University of British Columbia, 2005

A proposal submitted in partial fulfillment of
the requirements for the degree of

MASTER OF ARTS
in
INTERDISCIPLINARY STUDIES

We accept this proposal as
conforming to the required standard

.....
Dr. Brigitte Harris, Thesis Supervisor

.....
Dr. Wendy Schissel, Professor
School of Interdisciplinary Studies

.....
Dr. Matthew Heinz, Dean
Faculty of Social & Applied Sciences

ROYAL ROADS UNIVERSITY

December 2015

Abstract

To enhance TED Talk usability in ESL classrooms, I created a TED Talk corpus and determined a list of high-frequency vocabulary unique to the TED corpus. The investigation also determined the percentage of General Services List (GSL) and Academic Word List (AWL), New General Services List (NGSL) and New Academic Word List (NAWL) vocabulary and revealed an abundance of off-list vocabulary. Proceeding with the off-list vocabulary from the GSL and AWL, this exploration narrowed it down to 421 high-frequency word families and 2502 total word types, called the TED Word List (TWL). Finally, I assigned a TED Number to each TWL headword to assist in labeling the distribution across the corpus and determined the TWL vocabulary coverage across the TED corpus. The TWL can assist English language learners in determining how appropriate TED Talks are for study and help teachers determine what vocabulary may need special instruction.

Table of Contents

Abstract 2

Table of Contents 3

Acknowledgments 6

Chapter 1. Introduction 8

Research Questions 9

TED & Language Learning 10

Purpose of this Study 13

Significance of the Study 14

Study Limitations 16

Chapter 2: Literature Review 18

Second Language Acquisition. 18

English Language Teaching. 21

English for Academic Purposes 22

Corpus Linguistics 22

Corpus Linguistics and Language Learning 23

TED Talks as a Teaching Tool 24

Vocabulary Levels 31

Conclusion 32

Summary 33

Chapter 3: Research Design 34

Methodology 34

Methods and Analysis 36

Conclusion 47

Chapter 4: Results and Analysis 49

The Specialized TED Corpus 50

Are TED Talks Academic? 51

The GSL vs. the NSGL 52

Constructing the TWL 54

TED Number 59

TWL Coverage in the TED Corpus 61

Summary 62

Chapter 5: Conclusions and Future Directions 63

Key Findings 63

Recommendations 64

Implications of Research 65

Practical Applications of the TWL 66

Implications for Further Research 68

Conclusion 71

References 72

Appendix A 82

Definition of Terms 82

Appendix B 84

The TED Word List (TWL) 84

Acknowledgments

A huge debt of gratitude must to go to my wife, Tomoko Kajiki Wolfe, for continually supporting this long thesis journey. Now that this project is complete, I hope we can move on to one of your choice. I also want to thank my children Sora Wolfe and Wren Wolfe for not getting frustrated with me not playing with you as much as you (and I) would have liked. Time to catch up.

Thank you to Dr. Brigitte Harris for supervising me through the many ups and downs of this process, and to Dr. Rick Kool for his advice to move to Interdisciplinary Studies, and my official readers, Dr. Wendy Schissel and Dr. Matthew Heinze.

A big shout out of solidarity to my brothers and volunteer readers. Troy Hammond, for his editing and grammar check, and for being the person who introduced me to my first TED Talk. Matthew Allen, for editing and serious comma addition. You guys rock.

Thank you to TED speaker Sebastian Wernicke for the TED Excel file, Yuki Tanimoto for help with programming at the beginning of this investigation, and Leander Hughes for writing the giTED program to acquire the rest of the TED transcripts from the TED website (although I never got that thing to work right on my computer).

I would be remiss without a final thanks to the TED organization, for all the great videos and a special thanks to the TEDxTokyo community for all the life

lessons and opportunities to meet and assist people with their idea worth spreading.

This document is one of my ideas worth spreading. Enjoy.

Chapter 1. Introduction

If you are afraid to be wrong, you'll never come up with anything original.

-Sir Ken Robinson, Do schools kill creativity? TED Talk, 2013

Technology. Entertainment. Design. TED. Do you remember your first TED Talk? TED Talks are engrained in my work routine, as an English and science teacher, but also as a person, with my volunteer time and community outreach. The memory of the first one is quite vivid. One morning, a coworker came to the office and informed me that there was something he *had* to show me; we sat at the computer and watched Rives' "The 4 a.m. mystery" (Rives, 2007). *Are there more? I need to show these to the students*. These were my two comments because the TED Talk had done to me what I hoped to do to my students: inspire.

Over the years since, I have introduced TED Talks in most of my lessons, regardless of subject, and regularly watch TED Talks for personal pleasure. In fact, TED Talks and my personal pedagogy have become intertwined. I use TED Talks for listening comprehension practice and the transcripts for reading and general language study. I use them in content-area classes to supplement textbook materials and foster inquiry and critical thinking. I use them as good examples of how to give a presentation and make good PowerPoint slides. In one of my favorite lesson plans, I empower students to give their own TED Talk and share their own "idea worth spreading."

The classroom walls came down when I followed TED Talks into the real world. Five years ago, I became involved with yearly TEDx event planning in Tokyo where I guide Tokyo high school students in planning TEDxYouth@Tokyo, a 100% youth organized event. I also try to inspire my colleagues and fellow teachers with the education themed TEDxTokyoTeachers. To a lesser extent, I volunteer with the larger TEDxTokyo[1]. When it came to TED's position in the curriculum, especially as time went on and the popularity of TED Talks grew, I realized that I was far from alone in both this appreciation and curriculum integration of TED.

As an ESL teacher, I valued TED Talks as free web-based, authentic listening materials, especially given the apparatus of their multi-language subtitles to help comprehension. Although subtitles are useful, I wanted to further improve the use of TED Talks as ready-to-use material for ESL teaching and learning. Unfortunately, through my own use and conversations with other ESL teachers (Browne, personal communication, October 14, 2014), I have found that TED Talks, and specifically the vocabulary they encompass, are quite difficult and challenging for many language learners.

Research Questions

This study answers the following questions:

How could the creation and exploration of a TED Talk corpus enhance TED Talk usability in ESL classrooms?

[1] For more information on these three events, please see http://www.tedxtokyo.com/en/

Sub-Question 1: How academic, with regards to the percentage of Academic Word List (AWL) and New Academic Word List (NAWL) vocabulary, are TED Talks?

Sub-Question 2: What does a TED Word List (TWL), a list of high-frequency vocabulary unique to the TED corpus, look like?

Sub-Question 3: How could a TWL be used to improve ESL students' vocabulary comprehension of TED Talks?

TED & Language Learning

For learners of English as a Second Language (ESL)[2 *], TED Talks can be used as an easily accessible tool to watch as informal, authentic English listening practice, either at the students' or teachers' initiative. I presented on the usability of the TED website at a language learning conference in early 2013 (Wolfe, 2013), but I was years behind others who had been using TED for educational purposes. Popular education blogger Larry Ferlazzo (larryferlazzo.edublogs.org) compiled a series of websites dedicated to TED Talks for ESL classroom use in 2009, suggesting that fairly shortly after they went online, teachers were quick to use them. Over the years, more TED ESL lesson plans have surfaced. These have been offered on a more casual basis in places such as the dedicated website TEDxESL.com and the dedicated

[2] Although most of the literature differentiates between English as a Second Language (ESL) as the learning of English by non-English speakers in an English speaking country and English as a Foreign Language (EFL) as the learning of English by non-English speakers in a *non-English* speaking country, this thesis uses the term ESL as a blanket term to describe both ESL and EFL.

section of Lingua House (Lingua House, 2015). However, TED Talks are very difficult not just because they speak at authentic speeds and intonation, but more specifically, because of the vocabulary used in them (Browne, personal communication, October 14, 2014). It is common practice to assume a student needs to know approximately 98% of the words in a text, with a maximum of one in 50 words unknown, to comprehend it fluently (Nation, 2006; Hu & Nation, 2000; Waring & Nation, 1997). However, Laufer (1989) suggested 95%, although not ideal, is at least manageable.

When determining a vocabulary profile in a text, the General Services List* (GSL) (West, 1953) and the Academic Word List*[3] (AWL) (Coxhead, 2000) are often used. The GSL is a historical list of the 2000 most common words in English, and the AWL is a 570-word extension of the GSL which focusses on academic vocabulary. In TED Talks the GSL and AWL provide total vocabulary coverage around 90-92%. Looking at these percentages another way, up to 1 in 10 words are possible unknown words and leaves many students struggling to comprehend meaning of the vocabulary in the TED Talk.

This difficulty did not stop TED Talks from formally entering the ESL textbook market. In 2015 a partnership between Cengage Learning and National Geographic Learning used TED Talks in the *World English* 2nd Edition and the new *21st Century Reading* textbooks, each with a continuum of four books of increasing

[3] Words marked with an asterisk denotes words found in the glossary. The asterisk appears only on the first occurrence of the word, after the research questions.

difficulty in the series (More information at http://ngl.cengage.com/assets/html/ted/). Although I have yet to use the former series in a classroom setting, I have used *21st Century Reading* and have read through both series. World English is a more traditional ESL textbook that focuses equally on the four skills of reading, writing, listening and speaking. It uses some of the themes and biographical information about TED speakers and their speeches, and includes activities based on watching shortened TED Talks. *21st Century Reading* uses TED Talks as the driving force of each chapter and the content has a more academic and critical thinking focus, alongside the language learning. Both series solve the TED Talk vocabulary difficulty issue through the use of a scaffolding apparatus for vocabulary and a careful curating and truncating of the speeches. Scaffolding (Wood, Bruner & Ross, 1976), in line with Vygotsky's sociocultural learning theory (1978), refers to any assistance provided to learners that helps them attain the Zone of Proximal Development (ZPD) and thereby accomplish something that they otherwise would not be able to do on their own (Vygotsky, 1978). *21st Century Reading* provides a considerable amount of scaffolding, with language instruction and study in each chapter leading up to the shortened TED Talk. Scaffolding is effective, and although the shortened speeches are available online, it does make students dependent on those books. Since TED Talks are free, and a potentially good resource for all English language learners, could there be another approach to making the talks in their complete form more accessible?

Language learners benefit from high frequency language study (Nation, 2006) and the General Services List (GSL) and its complement, the Academic Word List (AWL) are two often used lists that, when combined, constitute about 90% vocabulary coverage in academic texts. Coxhead (2000) showed that the two lists show academic text coverage of 86% and Nation (2004) showed 85-91% coverage. How similar are the vocabulary profiles of TED Talks and other academic texts? Could the language coverage of TED Talks be increased by another list, one that is specific to TED Talks? One approach to solving the vocabulary difficulty issue would be to provide ESL students with another list of specific TED vocabulary to facilitate better understanding.

Purpose of this Study

Since creating listening activities for language learning was assumedly not part of the TED organization's original plan, and the Cengage partnership is still new, further study of TED Talks as a whole and their ESL efficacy is needed. TED Talks are given to audiences of professionals, although speakers generally avoid jargon. As a fan and an avid viewer with a critical eye to passing along recommendations to my students, I have observed the use of higher-level English. Dr. Charles Browne, one of the authors of the NGSL and NAWL, confirmed this high-level English use as well (personal communication, October 14, 2014).

Vocabulary lists are a popular way to organize "need to know" or essential vocabulary for ESL students and teachers. Vocabulary lists are by no means a novel idea. In 1953, without any electronic assistance, West published his 2000-word list of

high-frequency vocabulary, the General Services List (GSL). West's 10-year project and hard work was not without reward, as it remains a high-use list today, over 60 years later (McKeown, Beck, & Sandora, 2012). Coxhead (2000) further cemented the GSL's position as a key list was when she developed the Academic Word List (AWL), a list of 570 high-frequency words (not including any GSL terms) found in wide-ranging academic texts, as an extension of the GSL (Gardner & Davies, 2013). The AWL has become the *de facto* need-to-know vocabulary list for ESL students interested in study at an English university (Gardener & Davies, 2013) and it is still being used in many current ESL textbook series, including Cengage and National Geographic's *21st Century Reading*, Oxford's *Inside Reading* and *Q:Skills for Success*, to name only a few.

Significance of the Study

In this study I investigated a large corpus* of TED Talks to identify vocabulary that complements the still popular ESL word lists, the GSL and AWL. I compiled the TED specific vocabulary into a third word list. This third list, the Ted Word List (TWL), can assist both teachers and learners by enriching and identifying more need-to-know vocabulary for better comprehension of TED Talks as authentic listening materials. TED Talks are a free, popular resource that ESL teachers and students can easily take advantage of, but a comprehensive analysis of TED Talks for vocabulary has yet to be undertaken. Are teachers making the best use of TED Talks in terms of guiding students to the vocabulary they need to comprehend it? Can a third list help students with comprehension of TED Talks?

It is not ideal, nor recommended, to study vocabulary from a list. However, lists are beneficial for students to know what vocabulary is necessary for them to meet their learning goals. Not all students need the same vocabulary, and it is not a good idea or a wise use of time to try and learn and use every unknown word students come across in their studies. One way to guide students is, according to Nation (2008), to suggest students be proactive and selective in their vocabulary study and make different categories of new words. These categories could be, for instance: words that students need know well enough to read and write as well as use in oral communication; words that students need to recognize in listening and reading but not really speak or write often; and words specific to the students' interests or studies. I believe that the TWL would fit into the second or third category, assisting in recognizing some higher-level English and adding to understanding of future TED Talks. For example, teachers could assign TED Talks as homework, using the TWL to prime students with some of the vocabulary necessary to understand the language, as well as some of the culturally related topics that make up some of the TED Talks' subjects.

ESL students need to learn vocabulary, and Gardener and Davies (2013) state that this is a limiting factor in student success in tests and higher education. A third list of high-frequency words from TED Talks would complement the other lists, increase the total coverage, and engage students with new useful language. The TWL could be an additional tool in helping students and teachers navigate the complex language in the over 2000 TED Talks currently online.

In addition, I have not seen any linguistic or ESL studies that look at a majority of the TED corpus. There is an ever-growing number of TED Talks, because the website adds new talks every weekday, but the few studies published use only about 60 talks (Coxhead & Walls, 2012; Wang, 2012). This study looked at all the TED Talks from June 2006 to December 2014, or 1790 talks. Considering this large number of talks, the TWL as a learning tool is pertinent to both ESL students and teachers across many countries.

Study Limitations

The study is limited by the fact that the TED website is updated every weekday, and thus the corpus I am analyzing is perpetually growing and changing. The TED corpus, as long as I keep adding to it, is called a monitor corpus. A monitor corpus is a corpus that is continually growing and is usually used to determine patterns of changes in the specific text. Until TED stops hosting conferences, or at least until they stop posting talks online, the TED corpus is incomplete and thus the vocabulary analysis also incomplete and subject to change over time. However, given that TED Talks were first put online almost 10 years ago, it seems like an appropriate time to investigate some vocabulary trends over the past decade.

TED, and its style of presentations, is not without its limitations and vocal critics. There are the elitist charges leveled against the costs (Lacy, 2010); another common criticism is that the time limit, maximum 20 minutes, is not long enough to address the complex issues these talks explore (Morgan 2014; Bratton, 2013). Morgan (2014, para 7) further suggests that TED Talks are the reason that your next

keynote speech "will be 30 minutes, or 20, or even 15, not 60 (or 90 minutes – 90 minutes was typical a decade ago)" and is worrying because "you can't persuade people to change in 15 minutes … make them emotionally uncomfortable enough with the status quo to be ready to embrace something new." Moreover, Bratton (2013) pulls no punches when he claims that the promises of change and solution commonly espoused in TED Talks rarely come to pass due to oversimplification, technocratic bias, and giving the false impression that you can skip all the difficult parts to solution-making with fancy, edited rhetoric. Criticisms aside, the use of TED Talks as a tool for ESL students, perhaps with some caveat discussion in either the students' native language or English, is still beneficial.

Chapter 2: Literature Review

> The single story creates stereotypes, and the problem with stereotypes is not that they are untrue, but that they are incomplete. They make one story become the only story.
>
> -Chimamanda Ngozi Adichie, The danger of a single story, TED Talk, 2009

The focus of this study is to create and explore a corpus of TED Talk vocabulary in the hope of facilitating better TED Talk usability in ESL classrooms. The purpose of the study is to determine how academic TED Talks are and to create a list of high frequency vocabulary unique to the TED corpus. This high frequency TED Word List (TWL) could be used to improve ESL students' vocabulary comprehension of TED Talks and facilitate greater English language acquisition. This chapter covers English language teaching, English as a Second Language (ESL), and their relationship with corpus linguistics. It discusses how TED Talks can be used as a teaching tool for listening practice as authentic materials, and much of this section discusses vocabulary and how the research and design of other famous word lists affected the creation and outcome of the TED Word List (TWL).

Second Language Acquisition.

Second Language Acquisition (SLA) theories are theories as to how people learn a second language and how their learning be improved. There are many different theories and they have undergone many transitions over the last century. Across these theories, there are three characteristics that most linguistic theories agree

on. The first is languages are systematic. As Chomsky (1957) eloquently pointed out with his famous example, “Colorless green ideas sleep furiously”; there is no limit to the number of new sentences and combinations of words in any language, but there are grammatical, cultural, social, and other rules, that must be learned and followed. The second agreed upon characteristic is that languages are also symbolic, meaning the visualized words and auditory sounds invoke a particular meaning agreed upon in each language. Finally, language is social. Most linguists believe that people have a built-in ability for language, but without the social context to develop language acquisition, language not acquired (Seville-Troike, 2013).

The SLA theory most relevant to my study is Krashen’s five-hypothesis based Monitor Model (1978). The five hypotheses are the Acquisition/Learning hypothesis, the Monitor hypothesis, the Natural Order hypothesis, the Input Hypothesis, and the Affective Filter hypothesis. The first hypothesis, the Acquisition/Learning hypothesis, suggests that we acquire much more language than we learn, just by being around samples of the second language. Importantly, this happens without conscious direction and desire. The second, the Monitor hypothesis, claims that people draw from this ‘bank’ of acquired language when they take part in unprompted conversation. The third, the Natural Order hypothesis, states that there are predictable sequences, or a natural order, to second language acquisition, but “the language rules that are easiest to state (and thus to learn) are not necessarily the first to be acquired” (Lightbown & Spada, 2013, p. 106). The Input Hypothesis, also another name for the model itself, states that learners can only learn if there is “comprehensible input”, i.e.,

if the student understands the input, then acquisition can begin to take place. In a method that is quite similar to scaffolding and Vygotsky's Zone of Proximal Development (ZPD) (1978) mentioned in Chapter 1, Krashen uses the formula *i + 1* to denote that input should be just a little bit of a struggle, with *i* denoting the student's language already acquired and *1* is the words, grammar, pronunciation, etc., just beyond the student's level (Lightbown & Spada, 2013). Finally, the fifth, the Affective Filter hypothesis is somewhat of a caveat on the power of the comprehensible input, because it states that despite comprehensibility, students can still raise a mental barrier with negativity, boredom, stress, etc., and language not acquired if this barrier is up.

Although this model has been criticized as overly simplistic and impossible to test empirically (McLaughlin, 1987), it was extremely popular in the 1980s and 1990s (Seville-Troike, 2013; Lightbown & Spada, 2013). It has remained popular (Butzkamm & Caldwell, 2009) and current: both Yarahmadzehi, Ganji, and Mahdavi, (2015) and Behroozizad and Majidi (2015) used the model as the SLA grounding their research in different levels of captioning of videos for ESL learners. Krashen himself still maintains its functionality and suggests the theory's simplicity may be a key factor in its persistence (Korea Bridge, 2011). The currency of Krashen's theory is also demonstrated by the fact that I have yet to attend an ESL conference without hearing his name and/or the comprehensible input hypothesis come up at least once.

The idea of comprehensible input can be seen in many textbooks and ESL lessons that begin with activities designed to prime students for learning by making

the input—reading or listening—more comprehensible. These types of activities allow students to comprehend things with assistance (headings, figures, photos, graphs, etc.) that they would not be able to do on their own. Referred to as scaffolding, it is a very common technique in ESL textbooks. These tasks could include, but are not limited to, discussion questions on the topic, vocabulary exercises, photo descriptions, multiple choice, and fill-in-the-blanks. A linguistic investigation into TED Talks could help ESL students increase vocabulary awareness, and thus the comprehensibility of the content, and make the use of TED Talks as a study tool more effective. Instead of making only individual TED Talks comprehensible, as seen in the Cengage's *21st Century Reading* series (http://ngl.cengage.com/assets/html/ted/), is there something that could assist in the comprehensible input for all TED Talks?

English Language Teaching.

The academic world of English teaching is an acronym-intensive one and not without considerable overlap. Language learners have different needs and one way to acknowledge these needs is to divide up the ESL world into separate entities. According to Martin (2014), the broad category of English Language Teaching (ELT) is quickly parsed into native instruction and English as a Foreign Language* (EFL) or English as a Second Language (ESL). Studying English in Japan is an example of EFL because it is the learning of a language that is not spoken in the country or region of study. An example of ESL is found in a Japanese person in Canada studying English, where the target language is used all around the person (Nation, 2013; 2012).

EFL and ESL are further refined to English for Occupational Purposes (EOP) and English for Specific Purposes (ESP) and, and under ESP we can find English for Academic Purposes* (EAP). Although EAP can be further divided, EAP is the underlying direction of my research because TED Talks are often illustrative of complex and higher level English, and could be used to mimic university lectures.

English for Academic Purposes

EAP has some marked differences from general EFL teaching. Martin (2014) lists three of these differences: academic skills, prescribed lessons, and listening. First, EFL classes are language driven, whereas EAP classes are focused on the academic skills needed to function at a post-secondary level. Second, EAP classes are much more prescribed and teacher-driven with less 'on the go' discussions and investigations into questions from students. EAP classes that have more teacher-talk-time and less class differentiation with more focus on covering of essential material, mimic the style of many academic classrooms. Third, listening to a lecturer will, in most cases, dominate much of the ESL and EFL student's academic time in class (Martin 2014). TED Talks, as an EAP tool, can offer some further subject-based, authentic listening to assist in second language acquisition.

Corpus Linguistics

A corpus is a body of text, representative of a certain aspect of a language, and stored in digital format to be analyzed with software called a concordancer*. Corpus linguistics is a collection of methods for studying language via a corpus (Lancaster, 2014). It is also important that the corpus created matches the research

question it is built to examine (Lancaster 2014). In this investigation, I built the TED corpus to create a tool, a TED Word List (TWL) so the only texts in the corpus were the totality of TED Talk transcripts available at the time. As mentioned in the introduction, and discussed in detail in Chapter 3, corpus linguistics is used in this study as a methodology. The definition of corpus linguistics exactly is not agreed upon within the academic literature, but is divided into two distinct categories: corpus-based and corpus driven. McEnery and Hardie (2012) describe corpus-based studies, like this one, as studies that are used to explore a hypothesis or a research question in order to prove or disprove it. The other, corpus-driven studies, does not use the corpus as a methodology, but rather the corpus itself is the starting point, or the source of the hypothesis and theories about about language (McEnery & Hardie, 2012; Tognini-Bonelli, 2001). Chapter 3 describes my corpus-based study in more detail.

Corpus Linguistics and Language Learning

While the first digital corpus, the Brown Corpus, was introduced in 1967 (Kučera & Francis, 1967), it wasn't until over 30 years later that corpora were commonly used in language learning (McEnery, Xiao, & Tono, 2006). As noted by Leech (1997) at that time, and reiterated later by Keck (2004), there are three areas where corpus linguistics and language learning interact: Language descriptions from a corpus, language analysis of a corpus in the classroom, and learner corpora.

Language description from a corpus consists of patterns of language in the corpora that can be used in curriculum and teaching materials. There are many

dictionaries, textbooks, and workbooks available to teachers and students because of corpus linguistics, and Römer lists a large number of recent publications addressing this topic, but perhaps the most important point is that corpus linguistics and language teaching greatly influence each other (2008): "While [language teaching] profits from the resources, methods, and insights provided by [corpus linguistics], it also provides important impulses that are taken up in corpus linguistic research" (Römer, 2008, p. 112). The use of corpus linguistics for language description comprises the majority of its use for language learning (McEnery & Xiao, 2011).

The second, language analysis in the classroom, uses the corpora itself in the classroom as a learning tool; learner corpora, the third category, is collecting and analysis of language learners' data and analyzing it for patterns of errors, and other ways of improving the language learning of students. This study, like the two famous studies that preceded it to create the General Services List (GSL) (West, 1953) and the Academic Word List (AWL) (Coxhead, 2000), is an investigation in language description. More recently, the two New General Services List* (NGSL) and NEW-GSL were both created from very large corpora to facilitate English language learning (Browne, 2013; Browne, Culligan, & Phillips, 2014; Brezina, & Gablasova, 2013; New Academic Word List, 2014).

TED Talks as a Teaching Tool

As an ESL and subject teacher, I have found TED Talks to be a comfortable addition in a wide range of my lesson plans and curriculum. Similarly, I have met many educators who also incorporate TED Talks into their lessons, and as mentioned

in the introduction, TED Talks have found their way into popular culture. I think TED Talks make excellent ESL teaching aides and below I describe why they are beneficial for both the instructor and the student as listening tools, authentic materials, and vocabulary boosters.

Importance of fostering listening skills. The importance of listening in EFL classrooms is often overlooked or underdeveloped (Moore & Rante Carreon, 2012; Graham, 2009). Nunan calls listening the "Cinderella skill," one that is "overlooked by its elder sister – speaking" (2002, p. 238). Listening is something that is treated as a naturally acquired skill, and passive and secondary to speaking (Cheung, 2010), but a survey done by Ferris and Tagg (1996) recommends active, varied, and authentic listening be added to EAP courses to better prepare ESL students. Rost (2011) goes further and defines six types of listening that he recommends teachers should strive to apply in equal proportions: intensive, selective, interactive, extensive, responsive, and autonomous.

Intensive listening is concentrated on phonology, grammar, and vocabulary with the listener focusing on how and what is said. Selective listening occurs when listeners are trying to determine main ideas and gist, and perhaps the most common type of listening practice undertaken in classrooms today (Rost, 2011). Interactive listening is when listeners work together to determine meaning or elucidate missing information. In extensive listening, the learner listens to larger amounts and longer extracts to determine meaning. Extensive listening can be academic listening or listening for pleasure and the use of full TED Talks would fall into this category.

Responsive listening is opinioned, personal, and output- orientated. Responsive listening is about the learner's point of view, but the output is based on the input. Autonomous listening can be any of the aforementioned types, but occurs when the learner selects the listening and manages the progress and assessment. Looking at this exhaustive list of six types of listening, I note that all but autonomous listening are covered in many ESL textbooks and classrooms. TED Talks could certainly be used for all types of listening described by Rost (2011), and given the large number of TED Talks and wide variety of subject matter, they would also make excellent material for less profuse, autonomous listening.

Cheung (2010) suggests that teachers have not been specifically trained or prepared to teach listening, despite Feyton (1991) raising the importance of fostering listening skills in the EFL classroom over 25 years ago. Research explained by Nation (2012) shows that listening should actually be at the forefront of the EFL class, because it assists in the acquisition of other skills. Referring back to Krashen (1978), it only makes sense to me that a student cannot give a correct output without fully understanding the input. Not only teachers, but students who rate online listening activities as effective (Shao, 2012), feel that listening is their weakest skill; specifically, they cite difficulty dealing with the speed of lectures (Graham, 2006).

Listening to academic lectures is often difficult for ESL students due to the speed of the lecture or vocabulary, and students often use online lectures for listening practice (Smidt & Hegelheimer, 2004; Takaesu, 2013). The use of online lectures has been shown to improve the *incidental* learning of language, or learning without any

intent to learn (Smidt & Hegelheimer, 2004). I am not trying to suggest that authentic materials should replace textbook study, which would counter the listening advice of Rost (2011) above. Rather, authentic materials, defined and discussed in detail in the next section, should be used to augment learning and promote incidental learning alongside structured learning.

Authentic materials. TED Talks can provide ESL students with authentic listening material in addition to any of the more structured and polished listening activities found in textbooks. Authenticity is described, although not easily or without debate, as any text that has not been created specifically for language learning or simply, text designed for native speakers (Harmer, 1983; Nunan, 1989). The range of research into authentic materials involves many subjects beyond ESL such as "cognitive and social psychology, learner autonomy, information and communication technology (ICT), motivation research and materials development" (Gilmore, 2007, p 99). Gilmore (2007) goes on to list eight different definitions from the literature of what authentic materials are, suggesting that there is far from consensus on this issue. Like Gilmore (2007), I personally prefer the more inclusive definition that "an *authentic text* is a stretch of real language, produced by a real speaker or writer for a real audience and designed to convey a real message of some sort" (Morrow 1977, p 13). Summarizing much research, Lingzhu (2010) distinguishes authentic texts as having both different redundant and grammatical features, with the first describing the normal aspects of speech that are often associated with repetition, and the latter referring to the less polished, mistakes and all, linguistic structure. Lingzhu (2010)

also notes the different stylistic features, and different environmental features not found in textbook material.

Research has shown that authentic texts, in studies comparing them to ESL created texts, foster more linguistic improvement, including vocabulary (Porter & Roberts, 1981; Bacon, 1989; Duquette et al, 1989; Herron & Seay, 1991; Mousavi & Iravani, 2012). Hwang (2005) calls on educators and learners to embrace current and popular authentic materials to inspire and facilitate a more natural second language acquisition. Since TED Talks are given to an audience of English speakers, they are authentic and generally jargon free, and as mentioned earlier, very popular. They clearly fall into the authentic material category defined by Morrow (1977), and combined with the wide range of topics and speech lengths, TED Talks make excellent authentic materials. One aspect of authentic materials that often thwarts ESL students' efforts at tapping the vast reservoir of TED Talks, though, is vocabulary.

Vocabulary. Perhaps not surprisingly, vocabulary plays a very important role in both first and second language acquisition across all levels of education (Chall, 1996; Biemiller, 2010; Townsend & Collins, 2009; Vacca & Vacca, 1996; Schmitt et al., 2011). Gardener and Davies (2013) refer to vocabulary as the "single most important discriminator" in most high-stakes tests, including, but not limited to, the SAT, MCAT, GRE, and TOEFL (Gardener & Davies, 2013, p 1). Lists of "need to know" vocabulary are a common way of organizing subject-specific vocabulary, but a more general academic language is needed for ESL learners (Chung & Nation,

2003; Nation, 2008). Some of the more popular lists include West's (1953) The General Services List (GSL), Xue and Nation's (1984) University Word List (UWL), Coxhead's (2000) Academic Vocabulary List (AWL) and most recently Gardener and Davies' (2013) Academic Vocabulary List (AVL). The GSL was first published over 50 years ago (West, 1953) and as Coxhead (2000) mentions, the GSL has come under criticism for its size, age, and lack of revision. Despite this criticism, Coxhead (2000) went forward with supplementing the GSL with her own list, the AWL, in which she chose not to include any GSL words. This decision has, in turn, led to some criticism of the AWL. Gardener and Davies (2013) spend a good portion of their article justifying their list with guarded criticisms against the AWL, but acknowledge that the AWL went on to become the default and most widely used academic vocabulary list.

Some of the AWL's popularity can be seen in the numerous textbooks dedicated to or incorporating the AWL, including but not limited to, Huntley (2006), Savage and Mackey (2010), Schmitt and Schmitt (2011), and Zimmerman et al. (2012). It can also be seen in the academic publications that analyze corpora for the percentage of AWL words in numerous fields including medical research (Chen & Ge, 2007), finance (Li & Qian, 2010), applied linguistics (Vongpumivitch, Huang & Chang, 2009), agriculture (Martínez, Beck, & Panza, 2009) and even a small corpus of TED Talks by the author of the AWL (Coxhead & Walls, 2012). Across most of these studies the AWL makes up about 10% of a given corpus. However, in the TED Talk analysis, the AWL coverage was just under half that at 4%, suggesting that TED

Talks use less academic vocabulary than academic corpora and more language akin to that used in newspapers (Nation, 2008).

These findings are not surprising since people speak differently than they write, and TED Talks are not academic presentations. Determining the AWL percentage of the TED corpus would give the linguistic community clarity where these transcripts are in comparison to other texts, academic or not. However, with only approximately 3% of the TED corpus studied by both Coxhead and Walls (2012) and Wang (2012), we are left with several questions about whether some talks are better than others in terms of enhancing ESL academic vocabulary learning, and whether there are any patterns of vocabulary that could assist in ESL learning using the TED corpus as a study tool.

Vocabulary lists are one way of determining and isolating essential vocabulary for ESL students. In the middle of the last century, West (1953) published his list of high-frequency English words, the General Services List (GSL). The GSL remains popular, and can be found in part, in whole, or simply mentioned in many academic or classroom-based ESL publications (McKeown, Beck, & Sandora, 2012). When Coxhead (2000) developed the Academic Word List (AWL), she chose to publish it as an extension of the GSL and did not include any GSL terms (Gardner & Davies, 2013). In 2013, two major studies created new General Services Lists. Browne (2013) and Brezina and Gablasava (2013) released the aptly titled New General Services Lists; however, the former uses the acronym NGSL, while the latter uses "new-GSL." Both of these lists are attempts to bring the general vocabulary

needed for language learners into the 21st century, citing the common criticism that the original is outdated in terms of content and methodology, i.e., no computer use and source material that is over 60 years old. Brezina and Gablasova (2013) released their New GSL based on a massive 12-billion-word multiple corpus; Browne, Culligan and Phillips (2013) released their New GSL based on a 270-million-word subsection of the Cambridge English Corpus. Taking up the idea of the AWL, Browne (2013) also released the New Academic Vocabulary List, or NAWL, (2014, February 17th) as an academic list that complements the NGSL.

Learning vocabulary requires a varied approach with multiple methods of instruction and must include repetition in both input and student production (Nation, 2000). Takaesu's (2013) practice with TED Talks as extensive listening and Reinders and Cho's (2010) recent technology trends in out- of-class listening with mobile phones, would work as supplementary material for EFL or EAP classrooms. Supplementing academic vocabulary study with carefully selected TED Talks based on the TWL would then be a beneficial and welcome approach to vocabulary acquisition by both the teacher and the student.

Vocabulary Levels

How large a vocabulary is needed to listen to TED Talks and understand them? To quote Nation (2006), "how much unknown vocabulary can be tolerated in a text before it interferes with comprehension?" (p 61). Hu and Nation (2000) claim 98% coverage, or one unknown word in 50, is the amount of vocabulary coverage needed to adequately comprehend a text. Laufer (1989) suggests that 95% is the

amount needed for basic comprehension and Hu and Nation mentioned that some people in their study were able to comprehend at that level; however, it does not make comprehension easy, only possible (Carver, 1994). More recently, Nation (personal communication, November 21, 2015) stated that 95% would still be too difficult for learners of a language, and suggested that the higher the coverage the better for, especially for fluency.

Determining what percentages of the first and second 1000 words of the GSL and the AWL the TED corpus comprises would also determine what vocabulary is off-list. This off-list vocabulary could be fashioned into a third list of vocabulary that could assist ESL students with reaching the 98% threshold for fluency comprehension. The closer students can get to 98-99% they better they do with the material and the more motivated they stay (Nation, personal communication November 21, 2015). By analyzing high-frequency words across the TED corpus, this study aims to compile a list of vocabulary that brings ESL students closer to comprehension.

Conclusion

TED Talks have already entered ESL classrooms via teachers' choices or textbooks. The language of TED Talks' is difficult for ESL students, and a TED Talk's effectiveness can be augmented through further strategic study. In order to facilitate more comprehensible input (Krashen, 1978) of TED Talks with some vocabulary scaffolding, this investigation created a list of high-frequency vocabulary unique to the TED corpus. I designed the TWL as a supplement to both the General

Services List (GSL) and the Academic Word List (AWL). Combined with the GSL and AWL, the new list can bring the vocabulary coverage of the TED corpus up to a comprehension level that more ESL students can understand, thereby making TED Talks more accessible and allowing more learners to use TED Talks as a free ESL resource.

Summary

This chapter has shown that the most relevant SLA theory to this investigation is Krashen's Monitor Model and described some of the ways in which language education is classified. It describes the history of corpus linguistics and shows that corpus linguistics can benefit ESL studies and language learners. Use of TED Talks in the ESL classroom has been validated in the literature for facilitating listening and vocabulary acquisition via authentic materials. This chapter looks at the importance of vocabulary for ESL students as well as the impetus and methodology for the creation of the General Services List (West, 1953) and the Academic Word List (Coxhead, 2000). It concludes with a discussion of language levels and what level is needed for students to comprehend a text without much difficulty. The next chapter addresses the research design used to make the TED corpus and the TED Word List.

Chapter 3: Research Design

> The best way to accomplish serious design … is to be totally and completely unqualified for the job.
>
> -Paula Scher: Great design is serious, not solemn, TED Talk, 2009

Delving into the research question "How could the creation and exploration of a TED Talk corpus enhance TED Talk usability in ESL classrooms?", this chapter looks first at the methodology of corpus linguistics used to guide the research and then the preparing and building of the TED corpus. My analysis of the corpus led to my creation of a word list of high-frequency vocabulary unique to TED Talks: A TED Word List (TWL). The final outcome is a list of 421 word families, and 2,502 words in total. With a goal of improving ESL students' comprehension, I formatted the TWL alongside two popular vocabulary lists–the GSL and AWL, added it to the vocabulary profiler software, to determine how much additional vocabulary coverage the TWL provided.

Methodology

The methodology used in this investigation is corpus linguistics. Corpus linguistics is the study of a body of text for a specific entity, with the text and the entity being extremely variable. Bowker and Pearson (2002) refer to it as "a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria" (p. 9). Recent trends and growth in both computing and digitization mean easier access to larger and larger corpora[*]–electronic bodies of language collected for analysis. It was relatively easy, although time-consuming, to

create and use a TED corpus. A TED corpus could be analyzed with respect to conventional corpus linguistic methodology such as, the number of unique words (i.e. token*), the total number of words (i.e. type*), and concordances*, or the display of every instance of a search word preceded and followed by a desired number of words for context. It also allowed me to determine what high-frequency vocabulary is unique to the TED corpus, and not found on popular ESL vocabulary lists such as the General Services List (GSL), New General Services List (NGSL), Academic Word List (AWL), and New Academic Word List (NAWL).

However, simply stating that corpus linguistics is a methodology is in itself somewhat contentious. Corpus linguistics is an academic field divided, albeit rather cordially, into two distinct entities: corpus-based and corpus-driven. Corpus-based linguistics "use corpus data in order to explore a theory or hypothesis … in order to validate it, refute it or refine it," while corpus-driven linguistics "claims instead that the corpus *itself* should be the sole source of our hypotheses about language […] that the corpus itself embodies its own theory of language" (McEnery & Hardie, 2012, p. 6). In essence, by referring to corpus linguistics as a methodology, I have firmly placed this study in the first group.

Of the four main goals of corpus linguistic quantitative analysis listed by Johnson (2008), this study uses data reduction, trying to reduce the large amount of vocabulary into a manageable list. Following Biber and Jones (2009) in identifying the first step, the primary unit of this investigation was "the word". Using this goal and first step, a quantitative analysis using corpus linguistics allowed me to answer

my research questions by looking at a majority of the talks that TED puts online as a unit of over three million words, reducing a large list of words that are not on popular ESL lists, determining what percentages of those words are on those lists, and finally narrowing it down into a new, manageable, TED Word list, a list that can enhance TED Talk usability and ESL student learning.

Methods and Analysis

This investigation with corpus linguistics started with the creation of a corpus. Although a monstrous task only 25 years ago and still quite complex and difficult even 15 years ago (Coxhead, personal communication, January 25, 2014), modern media and technology has allowed for digital corpus creation with ease. The TED corpus I created consists of 1,790 TED Talk speech transcripts posted to the TED website between June 2006–the date the first TED Talks went online–and December 31, 2014. I chose that end date for no reason other than to close the window of data collection and move on with analysis because the TED website continues to post new talks every weekday (TED, 2015).

In late 2012, before I came up with the idea of making a TED Word List, I already believed TED Talks made excellent ESL learning tools, and had been using them liberally in my classes. After showing a talk by Sebastien Wernicke (December, 2010) to a few classes, I contacted Mr. Wernicke and asked him if I could see the corpus he created, and he graciously –and very promptly–shared his corpus with me. The information he sent me in an Excel file was more of a data set than a corpus. It contained the first 916 TED Talk transcripts with a lot of other information like

speaker name, year, event, title, TED identification number, etc., all separated into rows and columns. In order to analyze the transcripts, I had to first separate each transcript into an individual file and I did this separation with a macro[4]. A macro in Excel is like an algorithm, or a set of instructions used to do repetitive tasks, and is written in the Visual Basic for Applications (VBA) language used in the Microsoft Office suite of applications. The macro successfully saved each transcript as an individual text (txt.) file. A custom computer program–referred to as a web-scraper[5]–automatically downloaded approximately 700 further transcripts from the TED website and turned them into text files. The remainder of the TED Talks, approximately 100 transcripts, were downloaded manually, a relatively slow process of copying the transcripts from the website and pasting them individually into text files.

I took many more steps to clean up the transcripts and get them ready for analysis. I downloaded many of the speeches in Rich Text Format (RTF) and these had to be converted to plain text (viz. UTF-8 character encoding). RTF adds formatting and code that gets converted to text, and in the analysis adds a lot of nonsensical 'words' and 'phrases', distorting the data, and the software used for the corpus analysis specified UTF-8. Many speeches contained non-spoken text, identified as words in parentheses, such as audience interactions and sound effects (e.g. "laughter", "applause", "booming", "buzz", and "hiss"). I manually removed

[4] Special thanks to Yuki Tanimoto, who wrote the macro to automate this process, saving me the copying and pasting of each one individually.

[5] Special thanks to Leander Hughes, a researcher at Saitama University in Japan, for the giTED program and saving me even more copying and pasting.

these and this process was assisted by a very useful free software, TextMate (https://macromates.com/), which allows the user to find and replace words or text across multiple files. For example, when the text "(Applause)" found it could be replaced with nothing and all instances of that text were deleted from the entire corpus. As I skimmed the text files in the corpus, I located a new non-spoken piece of text from one transcript and was able to remove it from all future transcripts with a simple find-and-replace (with nothing) request. In addition, although much less common, there were also sections of some transcripts where text on the speaker's slides, although unspoken, were transcribed into the speech. I confirmed the text marked by square brackets as unspoken text by periodic checks that compared the video and the transcripts, and summarily deleted them all.

I decided to delete the non-spoken aspects of the corpus because I wanted the TED corpus to be a purely *spoken* corpus and felt these non-spoken aspects would cloud the data. For example, the "(Applause)" text mentioned above appeared at least once per talk, at end of each speech, but sometimes throughout the talk as well, leading to a very high frequency in the corpus. Since studying these words would not directly assist the learner with necessary vocabulary for using TED Talks as a study guide, and having these words included would distort the frequencies, I decided to remove all non-spoken text from the corpus.

After the removal of of all non-spoken aspects of the corpus, the 1,790 text files were loaded into the free corpus profiler software, AntWordProfiler (Anthony, 2014). Unfortunately, this software could not process such a large number of files.

After some email communication with the software developer, which included me giving him the files to attempt to process himself, I concluded it was impossible with the current version. Therefore, I combined the 1,790 text files into one single file using the Apple software, Automator, and was able to use that single text file with the entire TED corpus in the profiler.

AntWordProfiler automatically loads with three lemmatized* level list files, the first and second 1,000-word families of the GSL, and the word families of the AWL. As seen in Table 1, the word files have more types in them than just the number stated. The first 1,000

Table 1

File Statistics of GSL/AWL & NGSL/NAWL

File	Types	Headwords (Families)
GSL 1st 1000	4,114	998
GSL 2nd 1000	3,708	988
AWL 570	3,082	570
1st NGSL	3,197	1,000
2nd NGSL	2,971	1,000
3rd NGSL	2,313	801
NAWL	2,605	963
NGSL Supplemental	174	50

words of the GSL (actually 998) has 4,114 types of words in 998 word families (e.g. the word *a*, has the word *an* in its family, so together they constitute one headword* with two types). The settings used were the default plus the additional output settings boxes, *Word Types*, *Include complete frequency list*, and *Include words in user file(s) but not in level list(s)*. To compare the GSL and AWL to the NGSL and NAWL in the

TED corpus, I removed the preloaded files–along with the first and second 1,000-word families of the NGSL, the next 801-word families, the 957-word families of the NAWL, the NGSL supplemental list of 50-word families of number words, and day and month names, for a total a five files (see Table 1). I loaded these into AntWordProfiler and run with the same settings as earlier mentioned.

Table 2

AntWordProfiler Lexical Profile Statistics of the AWL/GSL in the TED corpus.

File	Token	Token (%)	Type	Type (%)
GSL 1st 1,000	3,229,765	83.49	3,839	7.56
GSL 2nd 1,000	175,554	4.54	3,270	6.44
AWL 570	144,306	3.73	2,610	5.14
Off List	318,765	8.24	41,042	80.85
TOTAL	3,868,390	100	50,761	100

Looking at Table 2 and Table 3, we can note the high number of off-list word types, 41,042 for the GSL/AWL and 41,145 for the NGSL/ NAWL. That finding required more scrutiny, but this type of analysis is not possible in AntWordProfiler. Therefore, I saved the output from AntWordProfiler in an Excel file–for both further manual examination and the data reduction mentioned at the beginning of this chapter. A 40,000-word list is not useful for ESL students and I had to reduce this list to a manageable size.

At this point, I decided to focus only on the GSL and AWL due to their staying ability as the default lists for ESL textbook use, despite fair criticisms of their reliability; more on this decision in the analysis section. Looking at the GSL and

AWL output file, I sorted the 41,042 words into alphabetical order and, manually, into word families. If the total word family had a 100-word-or-higher occurrences, it was kept for further analysis. The minimum frequency of 100 was the same frequency minimum that Coxhead (2000) uses to make the AWL from a nearly identically sized corpus. Once I completed the list of word families with 100 or more occurrences–about 430 words in total–I searched the word families individually in the corpus for variations of the family that included prefixes or compound words. I did this search by loading the original 1,790 files into concordance software from the same developer as AntWordProfiler, AntConc. See Figure 1 for a screenshot of an AntConc search of the *amaze* word family. These searches did change the total word count for a few words and served to verify my original, manual organization of the 40,000 plus words.

Table 3

AntWordProfiler Lexical Profile Statistics NGSL/NAWL

File	Token	Token (%)	Type	Type (%)
1st NGSL	3,236,317	83.66	2,954	5.82
2nd NGSL	197,299	5.1	2,689	5.3
3rd NGSL	85,517	2.21	1,971	3.88
NAWL	58,641	1.52	1,893	3.73
NGSL Supplemental	24,246	0.63	109	0.21
Off List	266,370	6.89	41,145	81.06
TOTAL	3,868,390	100	50,761	100

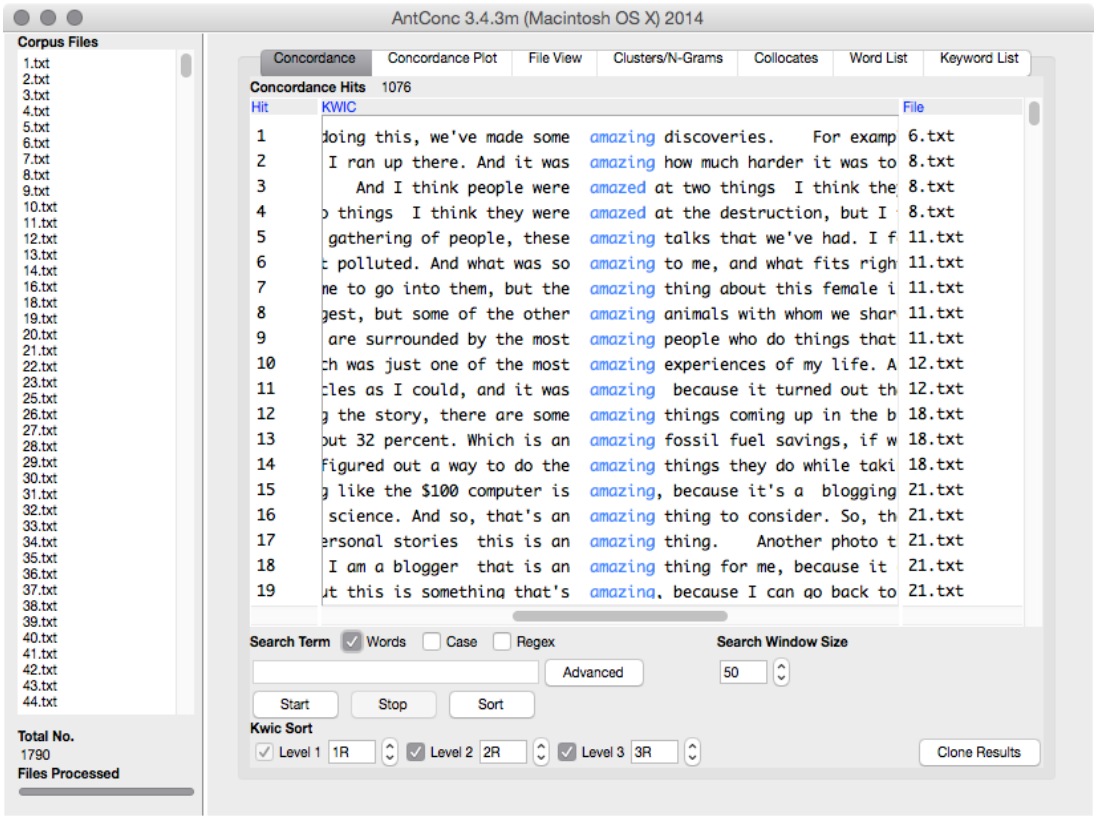

Figure 1: Screen shot of the concordance of the 'amaze' word family in AntConc. In this case the family had 1,076 'hits' in the 1,790 of the files in the TED corpus.

I also periodically eliminated headwords from the TWL. There were a few spelling differences that caused me to remove a few words from the TWL and put them back into the word families on the GSL or AWL. Some types of compound words (e.g. example *psychopharmacology*) were placed under both the headwords *psycho* and *pharmacy* and, since AntWordProfiler could not process double entries, they were placed into one family. I also removed the word types *et cetera* and *AIDS* from the preparation of the AntWordProfiler text file because two-word phrases and case sensitive searches cannot be processed by AntWordProfiler. As seen in Table 4, the final outcome of this quantitative reduction is a TED Word List of 421 high-

frequency headwords, or word families, and 2,502 types, including *et cetera* and *AIDS* (See Appendix B for the TWL headword list).

Table 4

AntWordProfiler File Statistics TWL

File	Types	Headwords (Families)
TWL 421	2,500	421

Note. The TWL actually includes 2,502 words, but *AIDS* and *et cetera* were removed before being run through AntWordProfiler. *Aids* (no caps) is an AWL word and case sensitivity and multi-word phrases are not processed by AntWordProfiler.

To determine the word family frequency, or the number of all occurrences of a word in the family across the TED corpus, I searched each word family as a group with AntConc. The results of one of these searches can be seen as "Total Plots" in Figure 2 below. I then divided the number of TED Talk transcripts the words occurred in by total occurrences to give what I refer to as a TED Number. The TED Number is between 0 and 1 and, as it approaches 1, this number indicates the degree to which the frequency in the word family is perfectly spread out at one occurrence per TED Talk. This perfect occurrence across the corpus is never attained but, as the TED Number approaches 0, the word may still have a high frequency, albeit in few talks, suggesting it has a more specialized usage. Table 5, below shows the 20 most frequent word families in the TWL, the total occurrences across the TED corpus, the number of TED transcripts the word appears in, and the TED Number.

Table 5

The first 20 most abundant word families in the TWL , number of TED transcripts found in, and TED Number.

Headword	Total	TED Transcripts	TED Number
ok	2,166	738	0.34
kid	2,113	578	0.27
cell	1,901	342	0.18
guy	1,824	674	0.37
gene	1,323	392	0.30
planet	1,270	570	0.45
video	1,085	488	0.45
amaze	1,076	576	0.54
incredible	1,005	175	0.17
cancer	989	258	0.26
robot	942	138	0.15
yeah	937	402	0.43
ted	931	439	0.47
internet	904	313	0.35
species	901	275	0.31
huge	851	514	0.60
drug	785	222	0.28
biology	733	285	0.39
math	719	240	0.33
laboratory	717	343	0.48

Note: As the TED Number approaches 1, it has a high frequency and coverage, occurring closer to once per talk. On the other hand, as it approaches 0, the word has a high frequency, but occurs in few talks, suggesting a more specialized usage.

Finally, I manually formatted and spaced the TED Word List (TWL) into the correct text file format for AntWordProfiler. Again, I did this formatting very slowly

and methodically in Excel by removing all the numerical cells, copying and pasting the headword in a new file, and then copying, and transposing the words in the family, so that they were vertical and one cell to the right of the headword.

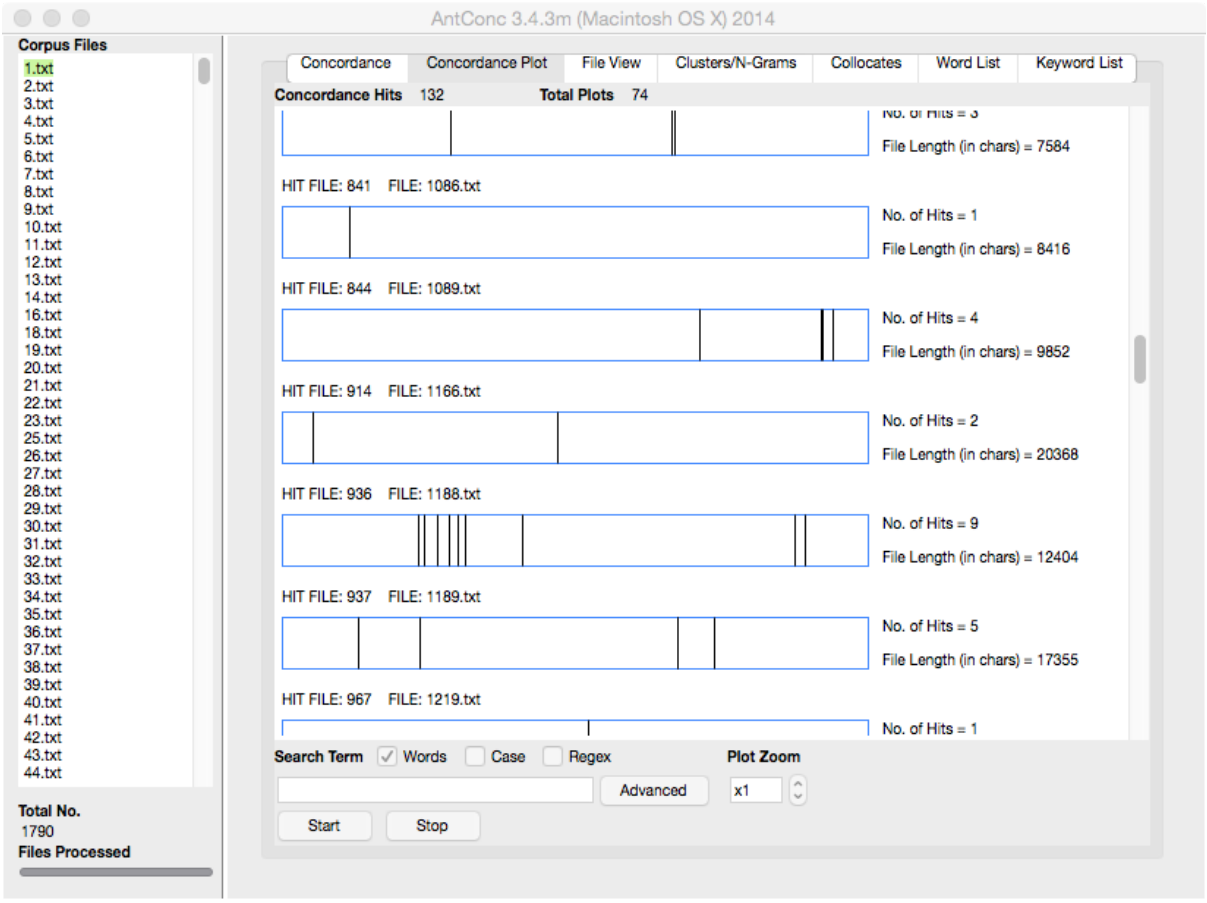

Figure 2: Screen shot of the concordance plot of the 'activism' word family in AntConc. In this case, the family had 132 'hits' in 74 plots (TED Talks) in the 1,790 files of the TED corpus.

I ran the word profiler analysis again with the two SGL and one AWL file, with the addition of the TWL file as a fourth file, to see how much further coverage the TWL provided in the corpus. The TWL provided an additional 2.7 % coverage reducing the off-list word from 6.9% (See Table 3) to 5.6% (see Table 6). At this time, it would have been desirable to do a similar analysis by adding the TWL list to the NSL/NAWL lists and running them through the profiler to compare coverage

between the two. However, since the TWL is based on the AWL and GSL, and not the NAWL and NGSL, and especially since the NGSL is based on a much more modern corpus than the GSL, there is a very large overlap of 774 word types. This overlap would not provide reliable data and would require a unique TWL based on the NGSL and NAWL; something which I considered, but which is beyond the scope of this investigation.

Table 6

AntWordProfiler Lexical Profile Statistics AWL/GSL/TWL

File	Token	Token (%)	Type	Type (%)
GSL 1st 1000	3,229,765	83.49	3,839	7.56
GSL 2nd 1000	175,554	4.54	3,270	6.44
AWL 570	144,306	3.73	2,610	5.14
TWL 421	104,135	2.69	2,500	4.93
Off List	214,614	5.55	38,540	75.93
TOTAL	3,868,390	100	50,761	100

Finally, I must mention the reliability, generalizability, and validity and of the methodology and design of this research. Hoover (2009) suggests that reliability is closely related to frequency in that words at the top of the TWL are more reliable that those at the bottom, but Kirk (2009) mentions that reliability is connected to corpus size, and the TED corpus is a little small. The TED corpus is small by today's standards, but one must remember that it is very representative of the specific genre it is looking at–because it includes all TED Talks up till the end of 2014. Biber and Jones (2009) also mention the reliability of computers, suggesting that this study may

have some issues because half of it analyzed by a computer and the other half human analyzed, and thus subject to human error. Generalizability, Biber and Jones (2009) continue, is linked to representativeness, or how well the corpus covers the desired aspect of the language. As mentioned above, this corpus is very representative, but–as Gavioli (2005) notes, the nature of the corpus is rather small and quite specialized; this reality reduces generalizability, but that result is to be expected. With the corpus and list derived only from TED Talks, it remains to be determined how useful the TWL is for other ESL activities. The validity of the TWL is defined by its representativeness of the list of high-frequency, useful vocabulary in the fist 1,791 TED Talks online. Validity cannot be determined based on frequency alone (McEnery, Xiao, & Tono, 2006), but the included TED Number allows users to determine how frequent the word is across the entire TED corpus, and thus determine validity for themselves. Also, some validity can also be seen in the large amount of overlap–774 tokens–with the more statistically rigorous NGSL and NAWL (New Academic Word List, 2014; Browne, 2013).

Conclusion

I compiled 1,790 TED Talk transcripts, removed all non-spoken aspects, and run the corpus through both profiler and concordance software. The output from that analysis showed that there was a large amount of off-list vocabulary: over 40,000 word types, or 6.9% of the corpus. Using a quantitative reduction, I reduced the 40,000 words to a list of 421 word families with 2,502 words in total. I added the new list to the GSL and AWL lists and ran them all through the profiler software again.

The total coverage that the TWL provided was 2.7%, reducing the off-list words by the same amount. Due to the significant amount of overlap between the TWL, NGSL, and NAWL it was not possible to compare the coverage between them. The next chapter analyzes and discusses the decisions made in making the TWL and what the list, and its 2.7% coverage, means for the ESL classroom.

Chapter 4: Results and Analysis

> There's zero correlation between being the best talker and having the best ideas.
>
> -Susan Cain, The power of introverts, TED Talk. 2012

The purpose of this investigation was to determine how the creation and exploration of a TED Talk corpus can enhance TED Talk usability in ESL classrooms. Specifically, the research questions were:

How could the creation and exploration of a TED Talk corpus enhance TED Talk usability in ESL classrooms?

Sub-Question 1: How academic, with regards to the percentage of AWL and NAWL words, are TED Talks?

Sub-Question 2: What does a TED Word List, a list of high frequency vocabulary unique to the TED corpus, look like?

Sub-Question 3: How could a TWL be used to improve ESL students' comprehension?

This chapter discusses the outcomes from the previous chapter, specifically, the creation and analysis of the TED Word List (TWL). It examines the corpus I used to make the TWL and the types of words I used in making decisions, addressing the research questions, justifying the use of the older but still popular GSL and AWL, explaining the TED Number, and evaluating the TWL by looking at its coverage across the corpus.

The Specialized TED Corpus

I determined directly from the AntWordProfiler output, some initial statistics and useful information. The total number words in the corpus, or tokens is 3,868,390. This corpus could be considered a relatively small corpus by today's standards (McEnery & Hardie, 2012), but it still represents a large majority of all the TED Talks posted online to date. In fact, Gavioli (2005) is hesitant to label corpora *small* or *large*, because the labels are misleading. For example, if a corpus contains 100% of a particular aspect or section of the desired language, to label it small is misleading. A corpus has to representative of the part of the language, meaning it represents well the aspect of language, writing (academic, fiction, email, etc.) or speaking (lecture, conversation, script, etc.), that is being investigated (Lancaster University, 2014; McEnery & Hardie 2012; McEnery, Xiao, & Tono, 2006). Instead of *small* or *large*, Gavioli (2005) prefers the terms, *specialized* or *general*. At almost 4 million words, the corpus is perhaps moving into the larger end of the specialized corpus, but since the only guideline for entry into the TED corpus was being a TED Talk available online, the specialized title is a good fit (Gavioli, 2005). Labeling the corpus as specialized allowed me to make a specialized word list with a few conventions that builders of a general list may not have: focusing on word families instead of lemmas*, including morphological words or prefixes, and combining different word families around themes. These decisions are addressed below.

Are TED Talks Academic?

I answered Sub-question 1 first, and quite quickly, in this study. How academic, with regards to the percentage of AWL words, are TED Talks? The coverage of the TED Talk corpus puts the GSL first 1000, the first 1000 words of the GSL and generally a large majority of the words in any English text, at 3,229,765 tokens or 83.49%, and the second 1000 words of the GSL at 175,554 tokens or 4.54%. The AWL has 144,306 tokens or 3.73%, and off list words, including proper nouns, at 318765 tokens or 8.24% (See Table 2). The percentage of AWL words in the TED Talk corpus is 3.73% and only slightly less than the 3.9% that Coxhead and Walls (2012) reported in their much smaller sample of 60 short TED Talks. This finding adds some credibility to their methodology and selection process. 3.73% also confirms the findings of Coxhead and Walls (2012) that TED Talks are not in the same category as academic articles or "not academic," if the definition of academic is an approximate 10% AWL coverage in the text (Coxhead, 2000). TED Talks are much closer to the coverage of newspapers in their AWL content, reported at approximately 4% (Coxhead, 2000; Chung & Nation, 2003). Confirming Coxhead and Walls' (2012) suggestion, this distribution of vocabulary in my study also suggests that TED Talks "contain more specialized and current vocabulary…as well as more everyday spoken language that is not reflected in the GSL/AWL" (p. 61). In this study that "more specialized and current vocabulary" is what I decided to analyze to see what words are common throughout the corpus and useful for ESL purposes.

The GSL vs. the NSGL

The off-list words, the total number of unique word types*, in the TED corpus is 50,761, and this number is a lot for a second language learner to attempt to comprehend. However, there were some words in this list that students should be familiar with, those of the GSL and AWL. Since the first 1000 words of the GSL make up approximately 85% of all English text of speech, most learners will already be familiar with many. Similarly, the AWL is still a very popular list used in ESL textbook creation and can be found in part or in whole in the appendix of many books. If the ESL learner or teacher is staying up to date in the ESL literature, or using Charles Browne, Brent Culligan, and Joe Phillips' (2014) *In Focus* series from Cambridge, he/she may now be familiar with the NGSL and NAWL. When those list's words are removed from the total, the data shows that off-list words, words not found on either the GSL/AWL or NGSL/NAWL, make up 41,042 and 41,145 respectively, or 8.24% and 6.89 %. Although the total number of words are almost the same, the lower percent of list words for the NGSL/NAWL validates Browne's (2013) claims that the NGSL/NAWL provides better coverage than the GSL/AWL. In addition, at the beginning of this investigation, I started making a TWL based on the NGSL/NAWL following the identical protocol described in Chapter 3. The TWL, based on 100 occurrences in the TED corpus, resulted in a list of only 233 word families. This outcome suggests that almost 200 headwords from the TWL based on the GSL/AWL are found in the updated and modern NGSL/NAWL. As mentioned in

the previous chapter, the 774 words types on the TWL that overlapped with the NGSL/NAWL further cements this notion.

This investigation started with a TWL based on the NGSL and NAWL, but I altered that plan and chose to work with the GSL and AWL and focus my list around those two still popular lists. It is because of a lack of dispersion and wide acceptance of the new lists at the time of this writing that I, somewhat reluctantly, continued with the less robust, but more prevalent list in the academic literature and used more widely in the ESL classroom and textbook, GSL and AWL. In 2013, two updates to the GSL, the New-GSL (Brezina & Gablasova, 2013) and the NGSL (Browne, 2013) and two updates to the AWL, the AVL (Gardner & Davies, 2013) and the NAWL (New Academic Word List, 2013) were released. Although predominantly better lists, because they address many, if not all, the problems underlining the GSL and AWL, they have yet to take hold in both the literature or ESL textbooks (except for *In Focus,* a series written by the NGSL authors). In Chapter 5, I comment on future directions, but one of the first things I wish to do when this research is complete, is finish the TWL based on the NGSL and NAWL.

So, working with only the GSL/AWL, I initially saved the word family if it had approximately 50 words in it. This decision was very arbitrary. Although the list I was reducing was in alphabetical order, prefixes and compound words would many times be unavailable until a thorough search with a concordancer. I had to use my own judgment to decide what words were in the family to be saved, and these judgments led to some reliability issues, so I proceeded initially with a conservative

approach. In keeping approximately 50 words, I ended up with an initial list of over 900 word families. A list of this size seemed too large to me, and referring back to Krashen's (1978) comprehensible input, I felt that a shorter list would be more comprehensible. In addition, I think that one of the reasons for the lasting popularity of the AWL is that it is a list of only 570 headwords and manageable for language learners as well as textbook publishers. For example, the excellent *Inside Reading* ESL books (Zimmerman et al., 2014) cover the AWL in five-level series. A list closer to 1000 words would be twice as hard to incorporate into classes and textbooks. Therefore, I doubled the minimum word family size to 100, the same number used by Coxhead (2000) in the AWL, and this size essentially halved the list to a more reasonable 453 words families. By comparison, the NAWL has 963 word families and Gardener and Davies' (2013) AVL is comprised of 3000 lemmas, where lemmas, similar to types, are only one form of the word.

Constructing the TWL

Through removals and combinations, I eventually narrowed down the 453 word families to 421 word families. I based these removals and combinations primarily on the literature's description of what types of words could be considered. McArthur (1999) identifies eight types of words that can be considered: orthographic, phonological, morphological, lexical, grammatical, onomastic, lexicographical, and statistical. Kirk (2009) adds 2 more: the numeral and discourse words. What follows is my explanation of which of the 10 types of words played a role in my investigation.

Orthographic words, words with dual spellings and dialect words, were not an issue in this study because transcripts were consistently in American English. However, some British spellings did remove a few words from the TWL because they were on either the AWL or GSL. One example, *advertise,* a headword from the GSL second 1,000 group, remained on the TWL. I kept the headword *ad,* and its family, *ads, adsense, advert, adverted, advertisers, advertizing, adverts*, on the TWL because of the high frequency of both *ad* and *ads*, and the many other variations of the word not included in the GSL word family (advertise, advertising, advertises, advertiser, advertised, advertisement, advertisements). This grouping seemed to me primarily an issue with the age of the GSL, because *advertisement* has commonly been shortened to *ad* over the years. I made other removals of some words that were off list only because that form of the word was not included in the original word family of the GSL or, less often, the AWL, for example *figured* and *figuring* (figure is in the GSL) and *tech*, *technologies, technologist*, etc. (technology is in the AWL). These were possibly not included in the original lists due to age of the list, limited usage, or change of use. I decided to remove these words despite having some new uses because anyone learning or teaching those word families would most likely be introduced, or at least come across these when studying. I did keep the words *app* and *activate* even though *application* is in the GSL and *reactivate* is in the AWL. The word *app* has become a whole new word with the advent of smartphones and reactivate may have been erroneously added to the *react* family in the AWL.

Morphological words, prefix morphemes such as eco-, bio-, and neuro-, were quite abundant in the corpus and kept. The morphological words included in the final TWL were nano-, neuro-, psych-, and ultra-. They were kept because even after words with those prefixes were added to other TWL families, the prefixes' totaled word count was still significant. This division was not the case with all the prefixes; for example, most words with the bio- prefix were placed with all the word families of the second part of the word on the TWL.

Lexical words, meanings within meanings and the context that comes with many words, although important, were not considered. Lexical meanings would look at slang, expressions, idioms, and sarcasm, areas that are not surprisingly difficult to understand for ESL students who focus more on understanding the words as they hear them. Due to the large number of individual TED speeches involved, as well as the large number of possible lexical interpretations across the many cultures that make up the TED Talk speakers, I did not undertake this investigation in this analysis. However, this study could be a future direction for using the TED corpus.

Onomastic, proper nouns, were included, because a few had a high enough frequency on my list to warrant inclusion, but the decision to keep some and remove others was very subjective. Countries and a large majority of people names were not included, but some famous people and company names were. Words like Facebook, Google, and Twitter, YouTube were kept because they are becoming more than just proper nouns, and actually have been used as verbs. More than that, they are a major component of recent culture and to ignore how they are used in modern speech about

current topics, both grammatically and socially, would be a disservice to language learners. I also kept the religious word families Christian, Jewish, and Muslim because these are some of the world's leading religions and also have quite a bit of controversy and conflict surrounding them. Again, these words were too important culturally and socially to ignore as a language component of study, and an ESL learner would be better served knowing them. I justified keeping these words by the fact that the TWL is a specialized list for TED Talks as a study tool, and unlike the GSL and AWL which are lists to help ESL students with all forms of English and academic English, respectively. Keeping them in the TWL suggests that these issues are common in TED Talks and the ESL leaner, or any TED viewer for that matter, should be somewhat familiar with these concepts.

Lexicographical, words with many different dictionary meanings (an example is the word *plant*, whose meanings can be green leafy organism; factory; put in the ground; hide; etc.) were also not considered when making the list, but should be when using the list for studying or teaching. This reality is in contrast to the other academic word lists mentioned at the beginning of this chapter that are lemma based lists, assumedly to give the learner a more specific direction with his/her studies. In the TWL, there are many word family members with only 1 or 2 occurrences across the corpus.

The statistical word, to "count each letter or group of letters preceded and followed by a white space," was the dominant word type for building this list (McArthur, 1999, p. 47). The statistical word is essentially a token, and was the first

piece of data collected from the TED corpus analysis as well as the first item discussed in this chapter.

Phonological words, words written with intended pronunciations, like *gonna* or *hafta,* or words written to include accents, were also not an issue because transcripts were not transcribed with them. Grammatical words, the differences between grammatical uses such as conjugations, were not specifically analyzed. Discourse words "can have a function which is neither referential nor propositional, which may also be pragmatically indeterminate, but which is, nevertheless, relevant for the development and cohesion" (Kirk, 2012, p. 31). Such words again were not an issue because the original transcripts did not make note of these. The original creators of the lists used to cross reference the TED corpus decided to include or not numeral words and were not further analyzed here because they did not meet the first criteria of appearing 100 times in the TED corpus.

A majority of the words in the TWL were combined based on word families and a theme, despite being different families. Some examples include, AIDS and HIV, Asperger's and autism, words related to Islam and Muslim, as well as tweet and Twitter, and words surrounding bees. These words were combined around a topic or theme, more than a headword connection. Again, I combined because of the specialized nature of the corpus and TWL to ease ESL students' understanding about a TED topic. In the future, if the TWL were to be the basis of a textbook, these words around a theme could make more engaging units with accompanying TED Talks.

In addressing the second sub-question, "What does a TED Word List, a list of high frequency vocabulary unique to the TED corpus, look like?", my analysis determined that 421 word families, just over 2,500 word types, with words in each family appearing in the TED corpus over 2,000 times in high end with *kid* and *ok*, and at least 100 times in the low end, with *addict*, *download*, *mushroom*, *pyramid*, and *scholar*. Since these word families do not appear in either the GSL or the AWL, the list is unique to the TED corpus and could act as supplemental tool to teachers and students using TED Talks as a learning tool.

TED Number

Since I had already determined simple word frequency, the next step was to determine how many TED Talks these words occurred in, or the general dispersion of each word across the corpus. Some words in the TWL are very unique to a certain topic and not ubiquitous throughout the corpus. I wanted to give users of the list a quick and simple way of knowing what words were unique and specialized and what were more general. To do this, I divided the number of TED Talks the words appeared in by the total number of the words in that family to get what I call a TED Number. The TED Number is between 0 and 1, and as it approaches 1, the word family approaches a theoretical 'perfect' dispersion of one word in the word family per talk it appears in. Given the very diverse subjects of TED Talks, it would be very useful to know how specific a TWL word is at a glance. The highest TED Number on the TWL was 0.84 with both the headwords *protest* and *ridicule.* More subject specific words would have a low TED Number and more general terms would have a

higher one. The idea of a TED Number came to me when I noticed that the word *password* appears 134 times in the TWL, but in only 13 talks, because it appears 119 times in one TED Talk about, not surprisingly, password security. The TED Number for *password* is 0.10, or 10%, and is the lowest on the list.

In this study I decided to forgo complex statistical analysis of the TWL and I made this decision for multiple reasons. The first was that I wanted this investigation to be grounded in usability and simplicity, and statistics can cause confusion for both the future user of the list and myself the researcher. I had considered Julliand et al.'s (1971) coefficient of dispersion, based on the standard deviation, but the results were not helpful, at least to this study. With a number between 0 and 1, the results put many words at 0, and according to Julliand et al.'s (1971) terminology, into one style group of the five they worked with: drama, fiction, essays, technical texts, and periodicals (Marie, 1992). This type of grouping was not helpful, nor were the values above 0, as I did not separate TED Talks and there was only one group in the TED corpus. Carroll's (1971) much more complicated index of dispersion, *D,* is also based on the assumption that the corpus is divided into sections, and again this division was not the case in the TED corpus.

The very specific subject matter of each TED Talk does make the TWL vocabulary disproportionate. Some vocabulary is very abundant, although its distribution across the corpus as a whole is very low, and this distribution is not a surprise given that one would expect different vocabulary use in talks about genetics compared to community development, or technology themed talks compared to

psychology. Despite this factor, I feel that the list is comprehensive, helpful, and combined with the TED Number, very usable. Future directions could include analysis based on themes, sex or nationality of the speaker, TED Talk ratings, etc., and the above mentioned statistics could be employed then.

TWL Coverage in the TED Corpus

Finally, sub-question 3 asked, "How could a TWL be used to improve ESL students' comprehension?" and to answer this question I looked at the TWL coverage within the TED corpus. The coverage that the first and second 1,000 words on the GSL, and the 570 words of the AWL is 83.5%, 4.5%, 3.7% respectively, and a total of 91.7%. Even rounding up to 92%, this number is still quite below the 98% recommended (Nation, 2006; Hu & Nation, 2000; Waring & Nation, 1997). However, when the TWL is combined with the previous lists, the 420 word families provide 2.7% more coverage, giving a total coverage of 94.4%. This number is very close to Laufer's (1989) minimum of 95% coverage needed for basic comprehension. However, in this study I did not determine the number of proper nouns in the TED corpus. Nation (2006) states that proper nouns are generally not very disruptive to language comprehension, and that proper nouns make up 4.55% to 6.12% of newspapers. Since the percent of AWL words in the TED corpus is almost identical to newspapers, we can assume that proper nouns make up a similar percentage in the TED corpus. With the very low estimate of 4% proper nouns in the TED corpus, and noted by Nation (2006) that proper nouns do not impede understanding much, the total coverage is now 98.4% and slightly above the 98% threshold for unassisted

understanding. The next chapter provides recommendations for use of the TWL in classrooms and individual study.

Summary

In summary, first I determined that the TED corpus I created was a specialized corpus, rather than referring to it as a small one. From that specialized corpus I determined that TED Talks are more aligned with newspapers than academic publications with regards to their academic content, i.e., the percentage of AWL vocabulary. Then, the I discussed difference between the GSL/AWL and NGSL/NAWL and the reasons for not moving forward with creating a TWL based on the latter. After that, I explained the types of words that were analyzed and used to construct the TWL, and why I gave each headword in the TWL a TED Number and what that number meant for usage. Finally, I determined, the percentage of TWL in the TED corpus and how much the TWL increased the total coverage of the vocabulary included on the GSL, AWL, and TWL. The coverage, not including proper nouns, increased the total coverage to 94.4%, an acceptable amount for unassisted comprehension. The section ended with some discussion about future uses of the TWL in the classroom and this discussion is continued in more detail in the following chapter.

Chapter 5: Conclusions and Future Directions

> When you look at the moon, you think, 'I'm really small. What are my problems?' It sets things into perspective. We should all look at the moon a bit more often.
>
> -Alain de Botton, Atheism 2.0, TED Talk, 2011

The broad focus of my study was the ESL learner and teacher communities. As an ESL teacher who uses TED Talks as leaning materials, I wanted to create a tool that would open, to other learners and teachers, more effective use of TED Talks to help meet their learning needs. More specifically, the focus was to build a TED Talk corpus, establish the vocabulary profile, and conclude how academic the TED Talk corpus is as measured by the percentage of Academic Word List (AWL) vocabulary. The motivation behind the vocabulary profile was to create a high-frequency vocabulary word list unique to the TED corpus.

Key Findings

The key findings of this study are both substantial and practical. The vocabulary profile, the percentage of General Services List (GSL) vocabulary, Academic Word List (AWL) vocabulary, and off-list vocabulary of the first 1790 TED Talks were 83.49%, 3.73%, and 8.24%, respectively. The percentage of AWL coverage, 3.73%, demonstrates that the TED corpus does not fall into the same category as publications in academic journals, which average approximately 10% (Li & Qian, 2010; Vongpumivitch, Huang, & Chang, 2009; Chen & Ge, 2007; Coxhead, 2000). Although this coverage sounds like good news for ESL students, i.e., they

don't have as much academic vocabulary to learn in order to understand TED Talks, there is a large amount of specialized vocabulary. In the study, I refer to that specialized vocabulary as off-list vocabulary and it appears in amounts that could easily thwart ESL students' efforts. This off-list vocabulary was over 8% of the TED corpus, or almost 40,000 word types; I refined them down into a list of 421 high-frequency headwords (Appendix B) and just over 2,500 total types. I called this list the TED Word List, or TWL. Alongside the TWL and its headword count, I determined how many TED Talks the word family appeared in as well as a TED Number. The TED Number gives TWL users a quick way to determine if a word has a wide distribution in the TED corpus or if it is more exclusive and relegated to fewer subject-specific talks. To determine the vocabulary coverage of the TWL in the TED corpus, I searched back in the corpus with the software AntConc, alongside the GSL and AWL, and established a new vocabulary profile. The TWL added 2.7% more coverage and with a low assumption of 4% proper nouns in the corpus, the TWL brings the TED corpus up past 98% total coverage including the GSL, and AWL. This amount is a noteworthy amount, because 98% is widely accepted as the amount of coverage needed for unassisted understanding (Nation, 2006; Nation, personal communication November 21, 2015).

Recommendations

For teachers, access to the TWL could assist in determining how appropriate TED Talks are in general, or how appropriate a specific TED Talk is for their students or class. Likewise, it would help teachers determine what vocabulary may

need special instruction and warrant some preliminary study. In a similar fashion, students studying independently can use the TWL as a guide in choosing whether or not TED Talks, again both as a whole or specifically, are appropriate for their vocabulary level and how much preparation is necessary. Like the GSL and AWL lists that proceeded it, the TWL and the TED Number can give students and teachers more precision and deciding whether or not a particular word is worthy of further study. When using TED Talks as a study tool, any unknown vocabulary that also appears on the TWL is undoubtedly worth noting, provided the student will study further with TED Talks.

Implications of Research

It must be reiterated that using the TWL assumes the ESL learner or teacher is aware of West's (1953) GSL and Coxhead's (2000) AWL; the TWL is an extension of those two lists and does not include any vocabulary from either. To approach the TWL without using the GSL or AWL would be to neglect much high-frequency and useful vocabulary.

Teachers and academics approaching this research now know where the TED corpus is in regards to academic vocabulary found on the AWL. Moreover, they have the additional knowledge that TED Talks also have a large amount of off-list vocabulary that could make their use in lower level classes more difficult. The TWL is one way to make the transition to TED talks a little easier by consolidating the large amount of vocabulary into 421 high-frequency word families. In the following section, I briefly outline some practical applications of the TWL.

Practical Applications of the TWL

Flashcards. The best way for a language learner to learn new vocabulary, upon first encounter, is to translate the word into the student's first language (Nation, 2001; 2008). After that, the word needs to be used to be remembered. TWL words can be put onto a flashcard, along with other useful information such as part of speech, pronunciation, example sentence, other members of the word family, etc. Although paper-based cards are still employed, app-based flashcards are another way to learn and monitor vocabulary study. There are numerous apps to choose from, but at the time of this writing, two free apps that are available on both the iOS and Android operating systems are Memrise and ZuKnow. Both of these apps allow for social sharing of flashcards and use spaced repetition, a popular, and in my opinion, much better method of vocabulary study. These social apps will also be helpful in the dissemination of the TWL, and is discussed in further detail below.

List. Teachers could provide students with a TWL list of headwords to use when reading or listening to TED Talks. When the students discover a word they do not know, they can cross-reference it with the GSL, AWL, and the TWL. If the word is on one of these lists, then this is a word that students need to recognize in listening and reading and need to know how to use in writing and speaking. If the word is on the TWL, then it is a word that students should recognize if they use TED Talks as a study tool.

Vocabulary profile. Predicting what might be difficult and challenging, and where students might falter or fail are some of the roles that teachers have to fill. The

ESL teacher must often predetermine what vocabulary and phrases students might have trouble with when a TED Talk is chosen for use in class. To assist in this identification, the teacher could employ the free software used in this study, AntWordProfiler, to find out what and how many words of the GSL, the AWL, and the TWL, are in the specific TED Talk he/she would like to use.

Vocabulary exercises. For learning new vocabulary and phrases there are numerous activities that could be employed: matching, fill-in-the-blank, true or false, choose the correct form of the word, etc. Using the TWL and TED Talk transcripts, I have begun creating some of these activities and hope to share them online once they are ready. The TWL is a starting point for my own curriculum design. I have started expanding out from the TWL into vocabulary activities and lesson plans for my students. Dependent on their popularity, these will be used as the basis for a future ESL textbook.

Dissemination of the TWL

I will publish the TWL and some of the findings of this investigation in academic linguistic and ESL journals. As seen with the numerous AWL percentage finding studies from different fields of academia, there is a desire to ascertain how much AWL vocabulary is in the literature. TED Talks are not academic but often are based on widespread academic subjects, are lower in subject specific jargon, and can often be very popular. Therefore, the academic profile of the AWL in the TED corpus would add to the linguistic literature, especially because TED Talk use in ESL is

growing. I will also apply to share the TWL and my findings at the Japanese Association for Language Teaching (JALT) academic conference in 2016.

Publishing in academic journals and speaking at academic conferences would put the TWL closer to researchers and academics, and possibly into classrooms, but would only be a first step. To promote the TWL, I need to make it available directly to ESL students. The easiest way is dissemination though various internet sites, such as the flashcard learning apps I mentioned above. This process would involve manually creating cards on multiple apps, allowing them to be downloaded, shared and used freely, and encouraging my own and colleagues' students to use them. Popular social media sites like Facebook and Google+ can also allow teachers and students to access the list. I will eventually make an interactive website for ESL teachers and students to use as a learning tool that would include not only the TWL and study materials, but also a vocabulary highlighter that would include many different lists, such as the NGSL and NAWL, the New-GSL, etc. Since TED Talk transcripts are available under Creative Commons (CC) free and fair use, there should be no issues with sharing the list broadly.

Implications for Further Research

During this study, I decided to move forward with making a TED specific word list as an extension of the GSL and AWL word lists, despite their known problems. As mentioned, this decision was due to the pervasiveness of the AWL across the ESL literature and textbooks. Since that decision, I have noticed that the New General Services List (NGSL) and the New Academic Word List* (NAWL) are

becoming more prevalent. The logical progression from this research would be to update the TED corpus with the few hundred new TED Talks available since its creation, and finish the NSGL- and NAWL-specific TWL. Based on initial data, that list would be about half the size, confirming Browne's assertion that the new lists have better, more modern, coverage.

As I mentioned in Chapter 4 when discussing the TWL creation, some types of words used for analysis were not employed in this investigation. Many of them would make for interesting analysis in future research. For example, lexical words, meanings within meanings and their contexts, would be of particular value for ESL students. When looking at concordances and word lists, I removed all subtleties and unspoken lexical meanings in the concordancing and word list creation during the building of the TWL. Consideration of how much TED speakers use slang, idioms, and sarcasm, would add another level of English language comprehension to the more often very literal attempts at understanding that ESL learners employ.

Since I made assumptions about the ratio of proper nouns in the TED corpus, determining the actual number of proper nouns would be worthwhile. My assumption was that TED Talks, having the same percentage of AWL words as newspapers and magazines, also have the same ratio of proper nouns. Nation (2006) noted that proper nouns make up 4.55% to 6.12% of newspapers, and that they are generally not a barrier to comprehension. Since I estimated the coverage of proper nouns at a very low 4%, this finding could actually increase the coverage of the combined GSL, AWL and TWL and strengthen the combined coverage of the three lists and the value

of the TWL as a study list. The second merit would be to document *what* proper nouns are abundant in the TED corpus. Further, because proper nouns are the names of people and places, proper nouns are great markers of popular culture and current events, and knowing what places and people are most abundant in the TED corpus would give ESL students some further knowledge about the culture that goes with the language. At this time, I am currently making a list of how many times each country appears in the TED corpus to determine what countries are omitted from TED Talks. Many countries have a very high frequency, but many are not mentioned even once. This information could be used to increase awareness of often overlooked countries, or give the next TED speaker an edge in his or her TED Talk proposal.

There are many areas of further research that could be explored in the future such as part-of-speech (POS) tagging, determining what type of word every token in the corpus is, and language differentiation between speakers, different nationalities, male and female, speech dates, etc., to see what differences and patterns arise. Other areas of further research could be with collocations, two- or three-word phrases, idioms, etc.

Since I made the corpus, TED has added approximately 250 new TED Talks to its online collection. One method of testing the resilience and usability of the current TWL would be to determine the TWL coverage in a corpus made with just the new TED talks. In order to update the current TWL, the TED transcripts would have to be added to the corpus and the same methodology followed. A TWL update would be much more time consuming than difficult, and the limiting factor is the number of

TED Talks released. Before embarking on revising the TWL, there should be enough TED Talks added to the corpus to make it worthwhile. After one or two years, and approximately 250-500 talks, there would be significant growth in the corpus, over 25%, and would signal a good time to update. Updating every two years would give some interesting patterns of language change as TED Talks are generally very recent in terms of their issues. It would be curious to know how and with what vocabulary the TWL is updated over the years to come.

Conclusion

The goal of this investigation was to make a usable tool for the ESL community of teachers and students. TED provides great, free online resources and the manageable-sized TWL, alongside the TED Number, can add some needed vocabulary in order to use TED Talks for classroom or independent study. If studied it can increase the vocabulary coverage for viewers or readers of TED Talks, increase English levels, and give people a chance to better their English language education and, hopefully, fulfill their study goals.

References

Anthony, L. (2014). AntWordProfiler (Version 1.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Bacon, S. M. (1989). Listening for real in the foreign-language classroom. *Foreign Language Annals, 22,* 543-551. http://dx.doi.org/10.1111/j.1944-9720.1989.tb02781.x

Behroozizad, S., & Majidi, S. (2015). The Effect of Different Modes of English Captioning on EFL learners' General Listening Comprehension: Full text Vs. Keyword Captions. *Advances in Language and Literary Studies*, *6*(4), 115-121.

Biber, D. & Jones, J. K. (2009). Quantitative methods in corpus linguistics. *Corpus linguistics: An international handbook*, *2*, 1286-1304.

Bowker, L. & Pearson, J. (2002). *Working with Specialized Language: A practical guide to using corpora*, London and New York: Routledge.

Browne, C. (2013). "The New General Service List: Celebrating 60 years of Vocabulary Learning". *The Language Teacher*, *37*, 13–16

Browne, C., Culligan, B., Phillips, J. (2014). *In Focus.* Cambridge University Press.

Brezina, V. & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, *36,* 1-22. doi: 10.1093/applin/amt018

Butzkamm, W. & Caldwell, J. A.W. (2009). *The Bilingual Reform: A Paradigm Shift in Foreign Language Teaching*. Tübingen: Narr.

Chung, T.M. & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language, 15*(2), 103-116.

Carroll, J. B. (1971). Statistical analysis of the corpus. In *The American Heritage Word Frequency Book.* Boston: Houghton Mifflin. xxi – xl.

Carver, R.P. (1994). Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Reading Behavior* 26(4), 413–437.

Coxhead, A. (2000). A new academic word list. *TESOL quarterly, 34*(2), 213-238.

Coxhead, A. (2011). The Academic Word List 10 years on: research and teaching implications, *TESOL Quarterly, 45*, 355–362.

Coxhead, A. & Walls, R. (2012). TED Talks, vocabulary, and listening for EAP. *TESOLANZ Journal*, 55-75.

Chen, Q. & Ge, G. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes, 26*, 502–514.

Chomsky, N. (1957). *Syntactic Structures*. The Hague/Paris: Mouton

Duquette, G., Dunnett, S., & Papalia, A. (1989). The effect of authentic materials in acquiring a second language. *Canadian Modern Language Review*, 43, 479-492.

Ferlazzo, L. (2009, June 3). The Best Teacher Resources For "TED Talks" (& Similar Presentations). Retrieved from

http://larryferlazzo.edublogs.org/2009/06/03/the-best-teacher-resources-for-ted-talks/

Ferris D. & Tagg, T. (1996). Academic Listening/Speaking Tasks for ESL Students: Problems, Suggestions, and Implications. *TESOL Quarterly, 30*, 297–320. DOI: 10.2307/3588145

Feyton, C. (1991). The power of listening ability: An overlooked dimension in language acquisition. The Modern Language Journal, 75, 173-180.

Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, 35, 305-327. doi: 10.1093/applin/amt015

Gavioli, L. (2005). *Exploring corpora for ESP learning* (Vol. 21). John Benjamins Publishing.

Graham, S. (2006). Listening Comprehension: The learner's perspective. *System, 34*, 165-182

Graham, S. (2009). [Review of the book Listening in the language classroom, by John Field]. System, 37, 540-541. doi:10.1016/j.system.2009.05.004

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching*, 40, 97-118. doi:10.1017/S026144480700414

Harmer, J. (1983). *The practice of English language teaching*: *New Edition*. London: Longman.

Herron, C. A., & Seay I. (1991). The effect of authentic oral texts on student listening comprehension in the foreign language classroom. *Foreign Language Annals*, 24, 487-495. http://dx.doi.org/10.1111/j.1944-9720.1991.tb00495.x

Hoover, D.L. (2009). Word Frequency, Statistical Stylistics and Authorship Attribution in Archer, D. (ed.) *What's in a Word List? Investigation word frequency and keyword extraction* 35-52. Ashgate.

Hu, M., & Nation, I.S.P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13, 403–430.

Huntley, H. (2006). *Essential Academic Vocabulary*. Cengage Learning.

Hwang, C. C. (2005). Effective EFL education through popular authentic materials. *Asian EFL Journal*, *7*(1), 90-101.

Keck, C. (2004). Corpus linguistics and language teaching research: bridging the gap. *Language Teaching Research*. 83-109.

Kirk, J. (2009) Word Frequency Use or Abuse? in Archer, D. (ed.) *What's in a Word List? Investigation word frequency and keyword extraction* (pp. 35-52). Ashgate.

Krashen, S. (1978). The Monitor Model for second- language acquisition. In R. C. Gingras (ed.), *Second Language Acquisition and Foreign Language Teaching* (pp. 1–26). Arlington, VA: Center for Applied Linguistics.

Korea Bridge. (2011, October 31*). Dr. Stephen Krashen Plenary KOTESOL International Conference 2011* [Video File]. Retrieved from https://www.youtube.com/watch?v=EXJwGFpfCY8

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Johnson, K. (2008). *Quantitative Methods in Linguistics.* Oxford: Blackwell.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren and M. Nordman (eds.), *Special Language: From Humans Thinking to Thinking Machines.* Clevedon: Multilingual Matters.

Lancaster University (Producer). (2014). *Corpus Linguistics: Method, Analysis, Interpretation* [MOOC]. Retrieved from: https://www.futurelearn.com/courses/corpus-linguistics

Leech, G. (1997), Teaching and language corpora: a convergence' in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.) *Teaching and Language Corpora*, 1-23. London: Longman.

Li, Y. & Qian, D. D. (2010). Profiling the Academic Word List (AWL) in a financial corpus. *System, 38*(3), 402–411.

Lightbown, P.M. & Spada, N. (2013). *How Languages Are Learned* 4th Ed. Oxford University Press.

Lingua House. (2015, June 10) EFL/ESL lesson plans: TED English. Retrieved from http://www.linguahouse.com/esl-lesson-plans/esl-course-plans/general-english-course-plans/ted-english/

Lingzhu, J. (2010). The Use of Authentic Materials in Teaching EFL Listening. *Humanising language teaching*, *12*(4).

Marie, T. (1992). *Quantitative Linguistics* (Vol. 37). John Benjamins Publishing.

Martin, P. (2014) 'Teachers in Transition: the road to EAP', in Breen, P. (ed.) Cases on teacher identity, diversity and cognition in higher education. Hershey: IGI-Global

Martínez, I.A., Beck, S.C., & Panza, C.B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study *English for Specific Purposes,* 28, 183–198.

McArthur, T. (1999). What is a Word? in T. McArthur (ed.), Living Words: Language, Lexicography and the Knowledge Revolution. Exeter: Exeter University Press, [1992]).

McEnery, T and Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). Corpus-based language studies: An advanced resource book. Taylor & Francis.

McEnery, T., & Xiao, R. (2011). What corpora can offer in language teaching and learning. *Handbook of research in second language teaching and learning*, *11*, 364-380.

McLaughlin, B. (1987). Theories of Second Language Learning. London: Edward Arnold.

McKeown, M. G., Beck, I. L., & Sandora, C. (2012). Direct and rich vocabulary instruction needs to start early. *Vocabulary instruction: research to practice*, 2, 17-33.

Moore, S.H., & Rante Carreon, J. (2012). Hidden challenges that radio DJs present to ESL/EFL listeners. *LEARN Journal: Language Education and Acquisition Research Network*, *5*, 19-29.

Morgan, N. (2014, July 10) What's The Problem With TED? Retrieved from http://www.forbes.com/sites/nickmorgan/2014/07/10/whats-the-problem-with-ted/

Morrow, K. (1977). Authentic texts and ESP. In Holden, S. (ed.). *English for specific purposes*. Modern English Publications, 13-15.

Nation, I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge University Press.

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59-82.

Nation, I.S.P. (2008) *Teaching Vocabulary: Strategies and Techniques*. Heinle Cengage Learning.

New Academic Word List. (2014, February 17th). New Academic Word List (NAWL) Homepage. Retrieved from: http://www.newacademicwordlist.org/

Nunan, D. (2002). Listening in language learning. In J. C. Richards & W. A. Renandya (Eds.), *Methodology in language teaching: An anthology of current practice*. 238-241. Cambridge, England: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511667190.032

Porter, D., & Roberts, J. (1981). Authentic listening activities. ELT Journal, 36, 37-47. http://dx.doi.org/10.1093/elt/36.1.37

Rives. (2007 March). Rives: The 4 a.m. mystery [Video file]. Retrieved from: https://www.ted.com/talks/rives_on_4_a_m

Römer, U. (2008). Corpora and language teaching. *Corpus linguistics: An international handbook*, *1*, 112-130.

Rost, M. (2011). Teaching and researching listening. (2nd ed.). Harlow, U.K.: Longman.

Savage, A. & Mackey, D. (2010). *Read This!* Cambridge: Cambridge University Press.

Smidt, E. & Hegelheimer, V. (2004). Effects of Online Academic Lectures on ESL Listening Comprehension, Incidental Vocabulary Acquisition, and Strategy Use. *Computer Assisted Language Learning, 17*(5), 517-556. DOI: 10.1080/0958822042000319692

Takaesu, A. (2013). Teaching Practice: TED Talks as an Extensive Listening Resource for EAP Students. *Language Education in Asia*, 4, 150-162. http://dx.doi.org/10.5746/LEiA/13/V4/I2/A05/Takaesu

TED. (2015, February 17th) TED Homepage. Retrieved from http://www.ted.com/

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

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.

Vygotsky, L. V. (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Wang, Y. (2012). An Exploration of Vocabulary Knowledge in English Short Talks: A Corpus-Driven Approach. *International Journal of English Linguistics, 2*(4), 33-43. DOI:10.5539/ijel.v2n4p33

Waring, R. and Nation, I.S.P. (1997). Vocabulary size, text coverage, and word lists. In *Vocabulary: Description, Acquisition and Pedagogy* N. Schmitt and M. McCarthy (eds.). Cambridge University Press, Cambridge: 6-19.

Wernicke, S. (Feb 2010). Lies dammed lies, and statistics (about TED Talks) [Video file]. Retrieved from: https://www.ted.com/talks/lies_damned_lies_and_statistics_about_tedtalks

West, M. (1953). *A General Service List of English Words*. Longman, Green.

Xue, G. & I. S. P. Nation. 1984. 'A university word list,' *Language Learning and Communication 3,* 215–29.

Yarahmadzehi, N., Ganji, M., & Mahdavi, M. (2015). The Effect of Audio-script Exposure on Students' Listening Comprehension in Dialogue Tasks. *Iranian Journal of English for Academic Purposes*, *1*(2).

Wolfe, J. (2013). TEDucation: Input and output; 2 ways of using TED talks in the language classroom. Proceedings from *The Asian Conference on Technology in the Classroom.* The International Academic Forum (IAFOR). Aichi, Japan.

Wood, D. J., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychiatry and Psychology*, 17(2), 89-100.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Zimmerman, C.B., Burgmeier, A., Zwier, L.J., Rubin, B. & Richmond, K. (2012). *Inside Reading: The Academic Word List in Context. 2nd Ed.* New York: Oxford University Press.

Appendix A

Definition of Terms

Academic Word List (AWL): The academic word list is a list of 570 words compiled after analyzing a corpus of magazines, newspapers, books, and journals and is the most occurring words across numerous disciplines.

Concordance: A display of every instance of a specified word or other search term in a corpus, together with a given amount of preceding and following context for each.

Concordancer: Software that can produce a concordance from a specified text or corpus.

Corpus (plural corpora): a corpus is a body of language representative of a particular variety of language or genre which is collected and stored in electronic form for analysis using concordance software.

English for Academic Purposes (EAP): English for academic purposes is the more specific training or learning of English for higher education and a university setting, and includes, but not limited to, proficiency in exams, presentations, essay writing, and reading.

English as a Foreign Language (EFL): English as a Foreign Language refers to the teaching of English in a non-English speaking area, such as Japan.

English as a Second Language (ESL): English as a Second Language is the study of English by people whose first language is not English. However, for lack of a better term, in this document ESL refers to both ESL and EFL.

General Services List (GSL): Created by West in the 1950s, it is a popular list of the first few thousand most popular English words.

Headword: Like a dictionary definition, a headword is a word that represents a list of words with similar meanings, for example the verb 'be' is the headword for am, is, are, were, being and been, and 'administrate,' for administer and administration.

Lemma**:** See Headword.

Lemmatization**:** The grouping of all word forms under the headword

New Academic Word List (NAWL): Released alongside his NGSL, Browne's (2013) NAWL is the modern update to Coxhead's (2000) AWL.

New General Services List (NGSL): In an attempt to modernize the GSL, Browne (2013) released a new version of this list based on a much larger and modern corpus.

Token**:** The number of individual words in a text or a corpus, often seen as the total word count.

Type**:** The number of unique words in a text with every word only counted once.

Appendix B

The TED Word List (TWL)

Below are the 421 TED Word List headwords in alphabetical order. The first number to the right of the headword is the total number of occurrences of that word family. The second number is number of TED Talk transcripts the word family occurs in. The third number is the TED Number; the total number of occurrences of the word family over the total number of TED Talk transcripts the word family occurs.

absorb	118	86	0.73	amaze	1076	576	0.54
abuse	185	100	0.54	amazon	146	64	0.44
accelerate	191	115	0.60	ancestor	197	94	0.48
accomplish	117	97	0.83	animate	204	91	0.45
acid	117	71	0.61	announce	157	126	0.80
activate	136	67	0.49	anonymity	109	48	0.44
activism	132	74	0.56	ant	280	40	0.14
ad	157	87	0.55	antibiotic	119	63	0.53
addict	100	61	0.61	apartment	139	71	0.51
ah	191	119	0.62	app	141	58	0.41
airport	108	75	0.69	architect	602	188	0.31
algorithm	231	77	0.33	asset	106	69	0.65
alien	139	81	0.58	asteroid	113	30	0.27

astronomy	126	51	0.40
athlete	106	46	0.43
atmosphere	264	117	0.44
atom	236	90	0.38
autism	210	33	0.16
automobile	103	56	0.54
awe	197	116	0.59
awful	161	117	0.73
bacteria	391	98	0.25
balloon	128	40	0.31
bang	160	78	0.49
barrier	105	79	0.75
bat	161	49	0.30
battery	230	102	0.44
beach	160	96	0.60
bee	369	49	0.13
beer	106	51	0.48
bet	129	93	0.72
bible	109	47	0.43
bike	182	75	0.41
biology	733	285	0.39
blog	271	91	0.34
bomb	290	131	0.45
boom	111	67	0.60
bored	188	136	0.72
brand	297	137	0.46
breast	196	57	0.29
brilliant	216	158	0.73
bubble	227	105	0.46
budget	191	118	0.62
bug	126	70	0.56
cable	112	53	0.47
campaign	253	130	0.51
cancer	989	258	0.26
capitalism	140	61	0.44
capture	267	176	0.66
carbon	445	150	0.34
career	323	202	0.63
cartoon	210	61	0.29
celebrate	285	183	0.64
cell	1901	342	0.18
CEO	112	72	0.64
chaos	127	79	0.62
charity	134	70	0.52
chase	108	82	0.76
chemistry	207	102	0.49
chimp	175	37	0.21
chip	190	80	0.42

Christian	138	66	0.48	cortex	158	55	0.35
chromosome	116	27	0.23	cosmic	140	61	0.44
circuit	183	75	0.41	craze	419	295	0.70
civic	365	166	0.45	crisis	369	172	0.47
classroom	216	94	0.44	cyber	141	34	0.24
click	213	114	0.54	Darwin	175	64	0.37
climate	540	190	0.35	dedicate	121	25	0.21
clinic	380	156	0.41	democracy	648	190	0.29
clip	172	108	0.63	demographic	102	54	0.53
cluster	109	49	0.45	dense	239	117	0.49
CO_2	183	80	0.44	deploy	119	75	0.63
cognition	212	109	0.51	desperate	111	91	0.82
collaborate	345	196	0.57	diabetes	102	49	0.48
column	105	48	0.46	diagnose	284	137	0.48
comic	110	43	0.39	diagram	110	65	0.59
compassion	420	78	0.19	diet	179	66	0.37
concrete	118	44	0.37	digital	511	220	0.43
confront	128	99	0.77	dinosaur	198	40	0.20
congress	179	88	0.49	disability	208	71	0.34
conserve	152	63	0.41	disaster	202	127	0.63
continent	226	113	0.50	disorder	201	69	0.34
coral	167	38	0.23	disrupt	106	57	0.54
correlate	104	64	0.62	DNA	539	144	0.27
corrupt	217	86	0.40	dolphin	152	27	0.18

donate	113	75	0.66	Facebook	217	108	0.50
download	100	76	0.76	fake	144	69	0.48
drill	136	68	0.50	fantastic	361	242	0.67
drone	121	84	0.69	fascinate	303	213	0.70
drug	785	222	0.28	feed	532	261	0.49
ecology	392	169	0.43	feedback	168	88	0.52
Einstein	148	57	0.39	feminine	111	42	0.38
elevate	115	77	0.67	fertile	131	77	0.59
elite	103	62	0.60	fiber	147	79	0.54
email	254	161	0.63	fiction	152	76	0.50
embarrass	118	73	0.62	filter	144	84	0.58
embed	120	71	0.59	flip	152	115	0.76
emergency	115	74	0.64	flu	203	54	0.27
emission	229	105	0.46	fluid	127	71	0.56
emotion	626	273	0.44	folk	206	123	0.60
empathy	176	87	0.49	footprint	104	59	0.57
empower	195	122	0.63	forever	206	158	0.77
engage	576	284	0.49	fossil	166	82	0.49
entrepreneur	257	104	0.40	frankly	123	94	0.76
epidemic	177	72	0.41	frog	127	54	0.43
era	182	110	0.60	frustrate	157	120	0.76
etc.	233	116	0.50	fuel	432	167	0.39
exponential	113	44	0.39	fusion	108	24	0.22
extinct	185	76	0.41	galaxy	328	56	0.17

gang	137	47	0.34
GDP	158	69	0.44
gene	1323	392	0.30
genius	104	70	0.67
geography	144	87	0.60
giant	258	172	0.67
glacier	105	32	0.30
glamour	127	16	0.13
google	448	206	0.46
GPS	112	46	0.41
grab	208	142	0.68
graduate	311	201	0.65
graph	354	184	0.52
gravity	187	85	0.45
greenhouse	109	50	0.46
grid	152	76	0.50
gulf	103	52	0.50
guy	1824	674	0.37
habitat	175	95	0.54
hack	238	62	0.26
height	155	112	0.72
hell	192	155	0.81
hero	306	160	0.52
hip	123	69	0.56
HIV	524	112	0.21
hormone	115	45	0.39
horrible	132	104	0.79
huge	851	514	0.60
humanity	355	217	0.61
humor	113	52	0.46
hydrogen	126	48	0.38
icon	131	91	0.69
illusion	142	67	0.47
immune	164	78	0.48
implant	101	37	0.37
impress	191	153	0.80
incredible	1005	175	0.17
infect	336	130	0.39
ingredient	126	62	0.49
inject	113	64	0.57
inspire	563	329	0.58
install	184	114	0.62
instinct	104	74	0.71
intellect	196	137	0.70
interface	195	98	0.50
internet	904	313	0.35
interview	241	123	0.51
intuit	231	126	0.55

invade	122	48	0.39	Mars	229	66	0.29
irony	117	73	0.62	massive	324	205	0.63
jail	128	68	0.53	mate	173	84	0.49
Jewish	164	71	0.43	math	719	240	0.33
journalism	230	106	0.46	mayor	122	54	0.44
kid	2113	578	0.27	meme	115	28	0.24
kilometer	201	106	0.53	mess	167	123	0.74
laboratory	717	343	0.48	metaphor	200	107	0.54
landscape	265	133	0.50	meter	290	153	0.53
laptop	191	83	0.43	microbe	146	31	0.21
laser	113	58	0.51	microscope	172	82	0.48
launch	333	198	0.59	mimic	120	99	0.83
lens	109	58	0.53	miracle	139	92	0.66
lifestyle	107	58	0.54	mirror	164	97	0.59
lion	123	47	0.38	mission	263	149	0.57
literally	549	325	0.59	mobile	349	169	0.48
liver	102	41	0.40	molecule	537	147	0.27
loop	128	71	0.55	mortal	199	95	0.48
magazine	235	150	0.64	mosquito	133	21	0.16
magic	424	213	0.50	movie	582	245	0.42
magnet	239	81	0.34	muscle	247	102	0.41
malaria	234	53	0.23	museum	316	135	0.43
mammal	145	64	0.44	mushroom	100	25	0.25
marine	159	68	0.43	Muslim	478	94	0.20

myth	151	79	0.52	patent	188	61	0.32
nano	195	63	0.32	peak	163	96	0.59
narrative	168	87	0.52	peer	179	99	0.55
NASA	137	75	0.55	penguin	147	16	0.11
navigate	112	79	0.71	pharmaceutical	118	88	0.75
negotiate	153	92	0.60	phrase	139	91	0.65
nerve	140	55	0.39	physician	143	58	0.41
nervous	161	98	0.61	physics	422	167	0.40
neuron	665	153	0.23	piano	123	87	0.71
novel	205	122	0.60	pill	106	44	0.42
nutrient	198	95	0.48	pilot	159	77	0.48
obese	144	50	0.35	pitch	127	66	0.52
obsess	181	113	0.62	planet	1270	570	0.45
ok	2166	738	0.34	plastic	337	125	0.37
online	480	213	0.44	platform	204	130	0.64
optimism	191	91	0.48	plot	124	82	0.66
orbit	153	53	0.35	plug	110	73	0.66
organic	187	101	0.54	polar	148	64	0.43
organism	347	108	0.31	pole	154	59	0.38
overwhelm	120	81	0.68	polio	105	17	0.16
oxygen	244	109	0.45	pollen	175	32	0.18
paralyze	128	59	0.46	pollute	196	96	0.49
passion	413	237	0.57	pop	216	159	0.74
password	134	13	0.10	Portrait	104	49	0.47

predator	131	62	0.47	rocket	173	81	0.47
pregnant	151	69	0.46	romance	149	72	0.48
privilege	145	112	0.77	rotate	132	74	0.56
professor	252	167	0.66	rural	164	101	0.62
profound	253	141	0.56	satellite	230	92	0.40
protein	314	101	0.32	scan	320	143	0.45
protest	120	101	0.84	scare	290	201	0.69
prototype	186	108	0.58	scholar	100	78	0.78
psycho	134	43	0.32	score	194	108	0.56
pyramid	100	48	0.48	script	190	71	0.37
quantum	133	47	0.35	sculpt	182	75	0.41
radiate	192	97	0.51	seal	101	47	0.47
rape	128	57	0.45	sensor	253	119	0.47
recycle	135	68	0.50	session	106	84	0.79
reef	124	34	0.27	shark	271	35	0.13
reform	166	68	0.41	skeleton	142	41	0.29
refuge	116	51	0.44	smart	460	269	0.58
remote	211	149	0.71	software	409	190	0.46
replica	172	91	0.53	solar	397	135	0.34
resilience	111	59	0.53	sophisticated	150	114	0.76
rhythm	138	56	0.41	species	901	275	0.31
ridicule	130	109	0.84	spectrum	142	78	0.55
ritual	103	42	0.41	spider	163	38	0.23
robot	942	138	0.15	spinal	115	50	0.43

stack	115	73	0.63	traffic	230	127	0.55
stare	117	93	0.79	tragedy	182	135	0.74
stimulate	178	99	0.56	transparent	204	105	0.51
storyteller	108	53	0.49	transplant	103	35	0.34
studio	147	87	0.59	tremendous	250	164	0.66
suicide	124	69	0.56	trillion	161	89	0.55
surgery	552	138	0.25	truck	190	107	0.56
switch	240	153	0.64	tumor	199	42	0.21
symmetric	190	48	0.25	TV	602	289	0.48
symptoms	140	65	0.46	twitter	273	102	0.37
synthesis	129	67	0.52	ultra	138	57	0.41
tackle	105	78	0.74	urban	306	137	0.45
tag	187	63	0.34	vaccine	274	55	0.20
talent	168	117	0.70	vast	230	170	0.74
tank	125	74	0.59	vegetable	165	94	0.57
TED	931	439	0.47	versus	182	128	0.70
teenage	190	121	0.64	vertical	107	73	0.68
telescope	167	49	0.29	vibrate	101	39	0.39
terror	365	130	0.36	victim	228	108	0.47
therapy	276	123	0.45	video	1085	488	0.45
tiny	478	280	0.59	virus	484	122	0.25
tissue	281	80	0.28	vital	108	79	0.73
toilet	135	55	0.41	vulnerable	227	118	0.52
toxic	170	80	0.47	web	467	174	0.37

website	326	194	0.60
weird	206	142	0.69
whale	204	50	0.25
wiki	167	59	0.35
wireless	119	32	0.27
worldwide	149	113	0.76
wow	183	132	0.72
yeah	937	402	0.43
YouTube	138	77	0.56
zone	198	124	0.63
zoom	139	83	0.60