A Corpus-based Investigation of Academic Vocabulary and Phrasal Verbs in Academic Spoken English

by

Hatem Aldohon

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Applied Linguistics and Discourse Studies

Carleton University
Ottawa, Ontario

Abstract

Academic spoken language is capturing the attention of many vocabulary researchers in recent years (Rodgers & Webb, 2016; Thompson, 2005). In turn, research has focused on assessing word lists to promote second language learners' comprehension of academic speech (Dang & Webb, 2014). The present thesis is comprised of two corpus-based studies investigating academic spoken discourse. The first study examined (1) the vocabulary demands of spoken academic English and (2) the coverage of Coxhead's (2000) Academic Word List (AWL) in spoken Academic English. Transcripts of 62 lectures and 7 seminars form the Michigan Corpus of Academic Spoken English (MICASE) were collected and analyzed. The findings suggested that coupled with proper nouns and marginal words, knowledge of the most frequent 3,000 or 7,000 word families is needed to reach 95.55% and 98.03% coverage respectively of the combined lectures and seminars corpus. The AWL provided 3.68% coverage of the combined seminars and lectures corpus.

The second study compared the lexical coverage of Garnier and Schmitt's (2015) phrasal verb list (PHaVE List) with the most frequent 150 AWL lemmatized verbs in academic spoken English. The analysis was carried out on a 2,431,351 running-word corpus created from the British Academic Spoken English (BASE) and the MICASE corpora. The finding indicated that the PHaVE List accounted for slightly higher coverage figures than the AWL in the study corpus. Pedagogical implications were made for English for academic purposes teachers, learners and material designers.

Acknowledgements

All glory and appreciation to the Almighty Allah- the most Beneficent and the most Merciful, Who enabled me to complete this thesis.

I owe my deepest gratitude to my supervisor at Carleton University Dr. Michael Rodgers for his invaluable academic support and thoughtful feedback guided me during the writing and revision process. This thesis would have not been possible without his professional guidance.

I would like to thank the examination committee members: Dr. David Wood (chair of defence), Dr. Sima Paribakht (external examiner) and Dr. Geoffrey Pinchbeck (internal examiner) for their kindness and insightful comments.

I would like to extend my appreciation to ALDS faculty at Carleton University for creating such a rich understanding of the discipline for me: Dr. David Wood, Dr. Jaffer Sheyholislami, Dr. Guillaume Gentil, and Dr. Eva Kartchava. I am also grateful to my program administrator, Joan Grant, for kindly answering my many questions while I was a student.

Heartfelt thanks to my amazing wife and children for being patient, understanding and supportive. I cannot thank you enough.

Dedication

I would like to dedicate this thesis to my mother, who has always been a driving force in my success.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Aims and Scope of the Thesis

The development of vocabulary knowledge is one of the most fundamental components of the second language (L2) learning process (Hirsh & Nation, 1992; Laufer, 1989; Milton, 2009; Nation & Webb, 2011; Rodgers, 2013). L2 research has identified the importance of individual as well as multiword vocabulary (e.g., *keep up with*) knowledge for language learners (Conklin & Schmitt, 2012; Hsueh-chao & Nation, 2000; Nation & Webb, 2011; Schmitt, 2004). However, the large number of English words presents significant challenges for the L2 teachers and learners due to limited class time dedicated specifically to vocabulary teaching. In order to cope with these challenges, many applied linguists tend to prioritize vocabulary teaching and learning based on frequency criteria (e.g., Nation, 2013; Nation & Beglar, 2007; Webb & Rodgers, 2009). A number of vocabulary researchers (see, for example, Coxhead, 2000; Garnier & Schmitt, 2015; Simpson-Vlach & Ellis, 2010; West, 1953) have made use of corpus[1] texts to generate single-item word lists as well as multiword items lists that represent the most frequent and salient vocabulary.

There are a number of academic word lists (e.g., Coxhead, 2000; Gardner & Davies, 2014; Xue & Nation, 1984). The General Service List (GSL) which was developed by West (1953) comprises a list of the 2000 high-frequency English words. The GSL has been useful over the decades, but with the advent of computer programs, a number of new lists have been created. One of the most famous of these lists is the Academic Word List (AWL) (Coxhead, 2000) which

[1] A corpus is a collection of sample texts that can be used to identify vocabulary words, patterns of words, or grammatical structures which are used more commonly rather than on the basis of chance alone (Nation, 2001).

has been widely viewed as representative of academic vocabulary in English written texts (Coxhead, 2011; Nation, 2006; Wang, Liang, & Ge, 2008; Ward, 2009). Academic vocabulary refers to those words that are prevalent in academic texts (Chung & Nation, 2004; Clark & Ishida, 2005; Coxhead, 2006; Nation, 2013; Schmitt, 2010), and thus they are different from frequent words which are usually used in everyday life or in a specific field of study (Nation, 2001).

Another more recent vocabulary list is Nation's (2013) BNC-COCA frequency list containing 29 word family groups. A word family is the headword, its inflections and common derivatives (Nation, 2013). The items of this list were derived from the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA). This list is often used to determine the frequency levels of words in English texts. The BNC-COCA corpora accounted for millions of written and spoken words. While the AWL and the BNC/COCA frequency lists are based on corpora and developed according to the concept of word families, they are different primarily in terms of function. The 570 AWL word families are often considered important for L2 learners to successfully manage academic written texts, whereas the BNC/COCA list are often used along with certain computer software to analyze other lists in corpora. The BNC-COCA list will be used in this thesis to investigate the frequency of the AWL words in academic spoken discourse.

As far as multiword items (e.g., *get through*) are concerned, the PHaVE List developed by Garnier and Schmitt (2015) contains a list of the most frequent phrasal verbs along with their most frequent meaning senses. A phrasal verb can be defined as a verb plus an invariable particle (e.g., *up, with, away*) "which functions with the verb as a single grammatical unit" (Quirk, Greenbaum, Leech, & Svartvik, 1985, p. 1150). Research for this thesis will be carried out with

the use of corpus linguistics tools and manual analysis to examine the distribution of the phrasal verbs contained in the PHaVE List in academic spoken English.

The scope of this thesis is rather broad as it deals with both individual words and multiword items. It primarily aims to explore the vocabulary demands of academic spoken English and to analyze both the AWL and the PHaVE List in academic speech in such a way that can help compare the value of these two lists for L2 learners. There are three reasons why this thesis focuses on both the AWL and the PHaVE List. First, the AWL as a single-item vocabulary word list is widely used in English for Academic Purposes (EAP) contexts as well as a reference for EAP researchers and developers of instructional materials (Schmitt, 2010). Second, concerning the PHaVE List, it is a pedagogical multiword verbs list including the most frequent phrasal verbs that were created according to their frequency of occurrence (Gardner & Davies, 2007; Garnier & Schmitt, 2015; Liu, 2011), and therefore, they are expected to be the most useful phrasal verbs for language learners to learn. In addition, no research has so far examined the distribution of this list in academic spoken English texts (see section 1.4 for a description of the rationale for studying this list). Third, earlier research has pointed out that language learners need both single words as well as multiword items to successfully understand different types of English texts (Conklin & Schmitt, 2008; Nation, 2008; Nation & Webb, 2011; Schmitt & McCarthy, 1997; Webb & Rodgers, 2009; Wilkins, 1972). For these three reasons, single words and multiword verbs fall within the scope of this thesis.

1.2 Why Academic Vocabulary?

In the EAP contexts, academic vocabulary is often described as "a layer of vocabulary that occurs across a range of academic subject areas" (Coxhead, 2016, p. 11), implying that learners may encounter academic words regardless of their field of study. A distinction is often

made by researchers between academic vocabulary and technical vocabulary which is used more frequently in a particular area of study, such as engineering and biology (Coxhead, 2000; Nation, 2013).

Vocabulary researchers largely agree that academic vocabulary is an indispensable part of L2 learners' education irrespective of their proficiency level (Gardner & Davies, 2013; Nagy, Herman, & Anderson, 1985; Nagy & Townsend, 2012; Nation, 2001; Schmitt & Meara, 1997; Webb, 2007). Nation (2013) posits several reasons for studying academic vocabulary. First, academic vocabulary is more common in different academic texts and less frequent in general texts. Second, L2 learners are often unfamiliar with academic vocabulary compared to technical vocabulary. Third, academic vocabulary typically accounts for a large number of items in academic discourse. Finally, academic vocabulary can be taught through explicit instruction, meaning that language learners can master academic vocabulary with some help and support from their teachers. Other researchers also point out that although learning academic vocabulary can be seen as a daunting task by learners of English as a foreign language (EFL) (Corson, 1997; Vongpumivitch, Huang, & Chang, 2009), it is necessary for college students to successfully master it if they wish to "be part of the academic community" (Coxhead, 2006, p. 3).

This study conducts a thorough investigation of academic vocabulary in academic spoken language and uses the Coxhead's (2000) AWL word list as a reference of academic vocabulary. As noted above, vocabulary knowledge also requires learners to be aware of both single words and multiword units (Pawley & Syder, 1983; Wood, 2015). The multiword items of particular importance in this thesis are phrasal verbs, which will be presented in the next section.

1.3 Why Phrasal Verbs?

Past research has indicated that formulaic language or multi-word expressions such as idioms (e.g., *kick the bucket*) collocations (e.g., *warm greetings*), and prefabs (e.g., *the thing is*) are widely used in the English language (Altenberg, 1998; Appel & Wood, 2016; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Conkilin & Schmitt, 2012; Folse, 2004; Martinez & Schmitt, 2012; Nation & Webb, 2011; Sinclair, 1991; Wood, 2015). Such multiword items comprise much of the English language (Willis, 2003), and therefore comprehending them becomes one of the essential and advanced aspects of word/ lexical knowledge in both spoken and written texts (Nation, 1990).

Phrasal verbs such as, *fall down, make up*, *put out* are one category of formulaic sequences. There is a wide acknowledgement among grammarians and applied linguists that phrasal verbs are of fundamental importance to EFL and English as a second language (ESL) learners (Celce-Murcia & Larsen Freeman, 1999; Courtney, 1983; Darwin & Gray, 1999; Gardner & Davies, 2007; Liu, 2011; Moon, 1998). The reasons for this importance are twofold. First, because of the high syntactic (e.g., separability) and semantic (e.g., literal versus idiomatic meaning) complexity of phrasal verbs, they are notoriously difficult for language learners to acquire, and as a result they often avoid using them (Bolinger, 1971; Celce-Murcia & Larsen Freeman, 1999; Courtney, 1983; Cowie & Mackin, 1993; Dagut & Laufer, 1985; Garnier & Schmit, 2015; Hulstijn & Marchena, 1989; Liao & Fukuya, 2004). Second, phrasal verbs are abundant and highly frequent in the English language. According to Li, Zhang, Niu, Jiang, and Srihari (2003), they account for one third of the English verb vocabulary.

There is a commonly held assumption among researchers suggesting that phrasal verbs occur more frequently in spoken than in written discourse (McCarthy & O'Dell, 2004; Side,

1990; Siyanova & Schmitt, 2007). This view is evidenced by corpus statistics indicating that the usage of phrasal verbs is greatest in conversation and informal spoken environments with over twice the frequency as in academic writing (Biber et al., 1999; Vilkaitė, 2016). These considerations support the idea that phrasal verbs are found in all different registers, from speech and informal writing to the most highly academic forms of English language (Cornell, 1985; Fletcher, 2005; Hsu, 2014), and therefore they are important for learners to gain mastery of the English language.

There are a number of corpus-based frequency studies that have explored the occurrence of phrasal verbs and provided important information about their frequency and use across written and general spoken discourse (Biber et al., 1999; Gardner & Davies, 2007; Garnier & Schmitt, 2015; Liu, 2011). However, to date no studies have looked at the (lexical) coverage of phrasal verbs on the PHaVE List in academic spoken language. Lexical coverage is defined as‘‘the percentage of running words in the text known by the reader’’ (Nation, 2006, p. 61). This study attempts to address this research gap by examining the presence of phrasal verbs in academic spoken English.

1.4 Why the PHaVE List?

The rationale for studying the PHaVE List is based on the research evidence indicating that phrasal verbs occur more frequently in spoken language (Celce-Murcia & Larsen Freeman 1999; Erman & Warren, 2000; Gardner & Davies, 2007; Liu, 2011; Moon, 1997; Vilkaitė, 216) and so it is expected that they would play a role in academic spoken language. In addition, because the AWL is based on academic written texts, it has been largely used to help L2 learners improve their vocabulary knowledge for academic written contexts (Nation, 2013). However, the AWL may not provide as much support for language learners in academic spoken settings, and

thus complementing AWL instruction to help with academic speech may be necessary. Finally, determining the lexical coverage of the PHaVE List would shed light on its potential value in academic spoken language.

1.5 Why Academic Spoken English?

English has become one of the major languages of academic spoken discourse. In fact, it is the most widely used medium of instruction for native and non-native English-speaking university students around the world (Hyland & Shaw, 2016; Long & Richards, 1985; Nesi, 2012). Although much of the literature investigating academic English has focused on written discourse (see, for example, Hood & Forey, 2005; Hsu, 2014; Hunston, 1994; Hyland, 1994; Laufer, 1889; Swales, 1990), academic spoken English is perhaps more relevant for university students and faculty. University students have to understand academic lectures, give academic presentations, participate effectively in seminars and classroom discussions, and most faculty and graduate students attend conferences before publishing articles in international journals.

The growing numbers of non-native English speakers joining higher-education institutions where English is used for instruction have aroused great interest among researchers in academic spoken English (Flowerdew, 1994; Rodgers & Webb, 2016). However, L2 university learners are known to have great difficulties in comprehension of spoken academic English (Cheng, Myles, & Curtis, 2004; Field, 2011; Flowerdew & Miller, 1996; Hyland, 2009; Kim, 2006; Lee, 2009; Rodgers & Webb, 2016). One of the most important factors for this difficulty is shortage of academic vocabulary (Kelly, 1991; Powers, 1985; Webb & Rodgers, 2009).

The research in academic spoken English has dealt with analysis of transcripts of various speech events gathered from academic corpora (Dang & Webb, 2014; Thompson, 2006; Webb &

Paribakht, 2015). Currently, there are a number of major English corpora available to researchers, such as the BNC and the COCA. These mega corpora involve a diverse set of situations and texts, but very few samples are accompanied with each type, and thus they may not be very helpful for the research of specific speech events or text types. This is the reason why more specific corpora have been produced. There are two specific academic corpora already available for research. One of these is the British Academic Spoken English (BASE) corpus consisting of 160 lectures and 40 seminars (totaling 1,644,942 words) and the other is the Michigan Corpus of Academic Spoken English (MICASE) containing 152 transcripts (totaling 1,848,364 words). The present study looks at the nature of vocabulary in spoken academic English represented in the BASE and MICASE corpora.

1.6 Organization of the Thesis

This thesis is made up of three chapters. Chapter 1 provides the rationale, aims and the key concepts of the thesis. The next two chapters are organised around two studies. In order to allow for more focused analysis, each study is introduced with separate literature review, research questions, methodology, findings, discussion, pedagogical implications as well as limitations and further research sections.

The two studies are concerned with academic spoken English. Study 1 (Chapter 2) looks at the lexical coverage of spoken academic English. It also shows the role of the AWL in potentially helping EAP learners to improve their comprehension of academic discourse. Study 2 (Chapter 3) explores the frequencies of phrasal verbs, as a type of formulaic language in academic spoken discourse. It focuses on the actual occurrences of the phrasal verbs included in Garnier and Schmitt's (2015) PHaVE List. This study also seeks to compare the presence of the PHaVE List and the most frequent AWL items in academic spoken English.

Chapter 2- Study 1: The Lexical Profile of University Lectures and Seminars

2. Literature Review

2.1 Introduction

This section begins with an overview of the categories of vocabulary. It presents the term of academic vocabulary and expounds upon the previous research on the distribution of the Coxhead's (2000) Academic World List (AWL) in English texts to justify the reasoning behind its use in this thesis. The second section shows the nature of vocabulary in both spoken and written texts as well as within academic spoken genres, such as lectures and seminars. The third section discusses the research conducted on the vocabulary thresholds that are necessary for language learners to infer unknown words from contexts and understand different English texts. This section is divided into three main subsections focusing on the relationship between vocabulary knowledge and comprehension of written language, the relationship between vocabulary knowledge and comprehension of general spoken language and the relationship between vocabulary knowledge and comprehension of academic spoken language. Finally, the last section reviews Dang and Webb's (2014) research which will offer background information with respect to the questions that will guide this study.

2.2 Different Categorisations of Vocabulary

2.2.1 Lexical and Functional Words

One categorization of English vocabulary is between content (lexical) words and functional words. Content words are those items that bear meaning and typically refer to nouns (*girl*), verbs (*occur*), adjectives (*amazing*) and adverbs (*frequently*), while articles (*the*),

prepositions (*on*), modals (*can*) and auxiliaries (*is*) belong to the class of functional or grammatical words (Hartmann & Stork, 1972; Quirk, Greenbaum, Leech, & Svartvik, 1985). Content and function words differ markedly in word frequency. Function words are typically more frequent and more widely used in different English registers than content words (Dworzynski, Howell, Au-Yeung, & Rommel, 2004; Howell & AuYeung, 2007). Schmitt (2010) maintains that one needs to go beyond the first frequent 100 word forms in a corpus such as the BNC to regularly encounter content words. Yet, it was found that the more genre-specific the corpus is (e.g., *finance corpus*), the more content words are found. For instance, Kennedy (2014) revealed that while content words accounted for 36% (18 words) of the most common 50 words in an economics sub-corpus, they accounted for only 6% (3 words) of the 50 most common words in a general academic English corpus, and accounted for only 2% (one word) of the most common 50 words in the whole sub-corpus (The Birmingham corpus).

2.2.2 High Frequency, Mid Frequency and Low Frequency Vocabulary

In relation to frequency criteria, Nation (2001) categorizes vocabulary into three types: high frequency, mid frequency and low frequency vocabulary. High frequency words may contain function vocabulary (*but, because, by*) and common content vocabulary (*college, provide*). High frequency words generally refer to the words at the 25th 1,000 and 2,000 BNC/COCA word level (Nation, 2012). Schmitt and Schmitt (2014) argue that the 3,000 word level should be included in the high-frequency level because these words are necessary to reach 95% coverage of tokens (running words) in most texts, and that the 3,000 word level represent the vocabulary that L2 learners need to know to effectively communicate in English. High frequency words (GSL) were found to account for coverage of 89% of purely spoken English (Nation, 2006), but account for a lower coverage of 80% of both spoken and written texts

(Nation, 2001; West, 1953). Mid frequency words represent the words between the 3,000 and 9,000 word levels in the BNC/COCA25 lists, and low frequency words refer to the words after 9,000 word level (Nation, 2013; Schmitt & Schmitt, 2014) with the 10,000 level comprising the more common low frequency vocabulary (Read, 2000). The reason for making the distinction between mid-frequency words and low-frequency words after 9,000 word threshold is that learners generally need high frequency words to gain sufficient comprehension (98% coverage) of most English texts (Nation & Anthony, 2013). An example of a frequency word list that combines high, mid, and low frequency vocabulary is Nation's (2012) BNC/COCA-25 word list (see section 2.11.2 for further description of this list).

2.2.3 General, Academic and Technical Vocabulary

On the basis of second language learning, Nation (2013) divides vocabulary into three main categories: general, academic and technical vocabulary. General vocabulary has been operationalized by Nation (2001) and West (1953) to be the 2000 high-frequency English words in the General Service List (GSL) and the first 3,000 word level in the combined BNC/COCA-25 list (Nation, 2012). It is believed that general vocabulary provides high coverage across all disciplines (Chung & Nation, 2004; Nation, 2001). Next, academic vocabulary, referred to as general academic vocabulary, is a group of words that are more frequent across different subjects of an academic nature (e.g., journal articles, textbooks, university lectures and seminars) but are less common in general texts such as informal social interactions (Chung & Nation, 2004; Li & Pemberton; Nation, 2001, 2013; Schmitt & Schmitt, 2005; Zhou, 2010). Finally, technical words are those common terms often found in a particular academic discipline, such as in anatomy (Chung & Nation, 2004; Nation, 2008), engineering (Ward, 2007) and health sciences (Lei & Liu, 2016).

Although these categories are generally useful in understanding how much vocabulary is needed, the distinction between them is not always clear (Muñoz, 2015; Nation, 2001). For example, some academic words can be classified as technical words based on their meaning in a particular context. Chujo and Utiyama (2006) found that the word *financial* which is a word in the AWL is classified as a technical word in a finance corpus. Another related debate refers to the distinction of academic, general and high frequency words. Hyland and Tse (2007) and Gardner and Davies (2013) claim that a small number of the AWL words overlap with the 2,000 most frequent English words (see section 2.2.6). Snow (2010), however, argues that:

> There is no exact boundary when defining academic language; it falls toward one end of a continuum (defined by formality of tone, complexity of content, and degree of impersonality of stance), with informal, casual, conversational language at the other extreme. (p. 450)

Of particular interest in this study is the academic vocabulary which will be examined in more details in the following section.

2.2.4 Academic Vocabulary

Coxhead and Nation (2001) suggest four characteristics that indicate the importance of academic English vocabulary. First, academic vocabulary is formal with a high frequency of occurrence across academic texts but is less frequent in non-academic contexts. Second, academic vocabulary makes up a considerable number of words in academic English texts. The 570 word families in the AWL cover around 10% of academic texts and 1.4% of the tokens in general English texts (fiction) (Coxhead, 2000). Third, the meaning of academic vocabulary is often unknown and can be challenging to EAP learners because they may encounter it in both

general and technical contexts in which vocabulary meanings may vary. Fourth, academic vocabulary is a type of vocabulary that teachers can explicitly help learners with.

Further research conducted on the usefulness of academic words confirms these arguments, indicating the supportive role of academic words in English academic texts (Campion & Elly, 1971; Cowan, 1974; Durrant, 2016; Farrell, 1990; Hutchinson & Waters, 1987; Nagy & Townsend, 2012; Nassaji, 2006; Praninskas, 1972; Staehr-Jensen, 2005; Trimble, 1985). For example, it has been found that academic vocabulary plays a role in L2 learners' academic achievements (Garcia, 1991; Li & Pemberton, 1994; Qian 1999, 2002; Snow, 2010; Stahl & Fairbanks, 1986), reading comprehension (Clark & Ishida, 2005; Park, 2012), writing skills (Cobb & Horst, 2015; Engber, 1995; Gass & Selinker, 2008; Gonzalez, 2013; Grobe, 1981; Shaw, 1991; Townsend & Kiernan, 2015) and academic English proficiency (Townsend, Filippini, Collins, & Biancarosa, 2012).

Despite the importance of academic vocabulary, it can be demanding to language learners (Rodgers & Webb, 2016). Nagy and Townsend (2012) claim that academic words involve several features that make them difficult for EAP learners to deal with. First, a large number of English academic words contain Latin and Greek origins (e.g., *acquire, economics*). Second, academic vocabulary is morphologically complex, implying that when prefixes and/or suffixes are added to a word they can change its meaning or grammatical function and create new words (for instance, the verb *employ* includes possible derivational and inflectional forms such as *employable, employability, employment, unemployed, unemployment, employee,* and *employer)*. Third, academic texts are lexically dense and academic vocabulary is laden with dense meanings. Lexical density is the percentage of content words (e.g., *verbs, nouns, adjectives, adverbs*) over the total number of words in a given text (Halliday, 1989). Fourth, academic words are more

abstract than general high frequency words (e.g., *metacognition*, *paradigm*). Further evidence to these claims comes from previous research that has pointed to EAP learners' difficulties with EAP reading (Clark & Ishida, 2005; Park, 2012; Vongpumivitch, Huang, & Chang, 2008) and writing skills (Li & Pemberton; Santos, 1988; Shaw, 1991; Townsend et al, 2012).

There is a further reason why academic words are difficult. Considering the large number of academic words found in a wide range of academic discourse, EAP learners may not have sufficient exposure to this kind of words, nor do they have sufficient time to spend learning them, and as a result, academic vocabulary becomes a demanding learning goal for these learners (Coady, 1997; Ferris, 2009; Gee, 1990). Accordingly, a number of vocabulary researchers have generally turned to corpus linguistics along with computer programs to create frequency-based lists that can help EAP learners prioritize their vocabulary learning (e.g., Coxhead, 2000; West, 1953; Xue & Nation, 1984). Laufer and Nation (1999) stress that a learner's first priority should be mastery of the English vocabulary that is highly frequent when they state that "words should be learned roughly in order of their frequency of occurrence, with high frequency words being the first" (p. 35). Although vocabulary frequency is not the only element that determines effective language learning, it is an important factor (Waring & Nation, 1997; Schmitt, 2010). There are two important single-item word lists directly relating to the present study, the GSL and the AWL. Before discussing these two lists, it is important to note that they were established on the basis of word families as counting units. Word families are those which "consist of a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately" (Bauer & Nation, 1993, p. 253). For example, the words *researched, unresearched, researchers, researches, researching,* and *researchable* are all belong to the word family *research*.

2.2.5 The General Service List (GSL)

Using a 5 million-word corpus, West (1953) developed the General Service List (GSL) containing the 2,000 general high-frequency word families in English in order to approach the needs of L2 learners. The language samples used to establish this list consist of textbooks, novels, essays, encyclopaedias, poetry, magazines and science books. The GSL has been used in the field of vocabulary research (McCarthy, 1990; Nation, 2001; Nation & Waring, 1997; Schmitt, 2000), and it also served as the baseline of the AWL in that Coxhead (2000) excluded the most frequent 2,000 word families, as defined in the GSL. According to Hu and Nation (2000), learners need to master these 2,000 English words since between 75% and 90% of any English text consists of them.

However, research done by Liu Na and Nation (as cited in Nation & Waring, 1997) showed that knowledge of the 2000 words only (between 75% - 90%) is not sufficient for L2 learners' overall comprehension of English texts. They argued that an L2 learner needs to know at least 95% of the words in a certain text to gain a good comprehension. Moreover, GSL has been criticized for being outdated (based on texts published before 1930's) and provided low coverage of high frequent vocabulary words appearing in academic discourse (Browne, 2014; Carter, 2012; Dang & Webb, 2014; Engels, 1968; Gardner & Davies, 2014; Nation & Webb, 2011; Richards, 1974; Schmitt, 2010).

2.2.6 The Academic Word List (AWL)

Drawing on the academic University Word List (UWL) (Xue & Nation, 1984), Coxhead (2000) developed a more representative academic list based on corpus principles. First, the text samples she selected are representative of a wide range of academic disciplines, including university textbooks, journals, and laboratory manuals found in the Wellington Corpus of

Written English (Bauer, 1993). Second, the entire corpus was categorized into four main academic domains (Art, Business, Law, and Science) consisting of 28 subject areas and having an equal number of tokens. To effectively identify the range of academic vocabulary that is frequent across all these domains, each of them was further divided into seven disciplines. For example, Art was subdivided into: Education, Linguistics, History, Politics, Philosophy, Sociology and Psychology. Finally, Coxhead (2000) considered the size of the corpus. The AWL was derived from an academic written corpus of approximately 3.5 million words with roughly 875,000 running words in each of the four main academic disciplines: Art, Business, Law, and Science.

In order to ensure that the AWL is generalizable to many different disciplines, Coxhead (2000) then made it taking into account several criteria for inclusion 1) the researcher excluded the 2,000 most frequent word families included in West's (1953) GSL, making a distinction between general high frequency vocabulary and technical vocabulary. Coxhead's rationale for this is based on the assumption that learners would be familiar with the GSL words, (2) the AWL word families were selected based on a wide range. A word family had to occur "10 times in each of the four main sections of the corpus and in 15 or more of the 28 subject areas" (Coxhead, 2000, p. 221) and 3) frequency criterion was also considered. A member of a word family "had to occur at least 100 times in the Academic Corpus" (p. 226) with an average of 25 times in every main sub-corpus (arts, commerce, law, and science) (Coxhead, 2000). The result was an academic word list of 570 word families. Coxhead (2000) highlights the instructional value of the AWL, suggesting that it "might be used to set vocabulary goals for EAP courses, construct relevant teaching materials, and help students focus on useful vocabulary items" (p. 227).

Although the AWL is based on a principled research design, it is not without shortcomings. For example, the AWL has been criticised for being based around the words found in the GSL which is viewed as outdated source (Gardner & Davies, 2013). Several researchers indicate that the GSL includes many words that are no longer in general use (e.g., *telegraph*, and *footman*) and excludes some newer English vocabulary that might be salient and high frequent (e.g., *television*, *computer*, and *Internet*) (Brezina & Gablasova, 2013; Cobb, 2010; Eldridge, 2008; Hancioğlu, Neufeld, & Eldridge, 2008; Nation & Hwang, 1995; Neufeld, Hancioğlu, & Eldridge, 2011). Indeed, Gardner and Davies (2013) examined the distribution of the AWL in the COCA and discovered that 79% of the AWL word families (451 of the 570) are within the top 4,000 words of the COCA, suggesting that a considerable number of the AWL words are general high frequency words of English.

Some researchers also question the usefulness of the AWL as it makes a distinction between academic and disciplinary vocabulary (Durrant, 2014; Hyland & Tse, 2007; Muñoz, 2015; Paribakht & Webb, 2016). Hyland and Tse (2007) argue that a number of the AWL words overlap with the 2,000 most frequent English words in the GSL, implying that the AWL may not fully represent academic vocabulary and that discipline specific vocabulary can be more helpful to the learners as it covers a large proportion of discourse. They criticized the assumption that "there is a general vocabulary of value to all students preparing for, or engaged in, university study" (p. 248). However, other studies on the coverage of academic words in discipline specific corpora (Chung & Nation, 2004; Li & Qian, 2010; Martinez, Beck & Panza, 2009; Vongpumivitch, Huang, & Chang, 2009; Ward, 2009) indicate the supportive role of the AWL in academic texts.

Although the AWL has been criticised especially for being not representative of general academic vocabulary (Brezina & Gablasova, 2013; Gardner & Davies, 2013; Hyland & Tse, 2007), it is still widely used in the teaching and testing of EAP (Nation, 2008; Schmitt, Schmitt, & Clapham, 2001), stimulated the creation of other more specialized academic word lists (Chen & Ge, 2007; Nelson, 2000; Wang, Liang, & Ge, 2008) and highly recommended by other vocabulary researchers (Coxhead, 2012; Nation, 2013; Schmitt, 2010). Therefore the AWL will be used as a reference to academic vocabulary in this study.

2.3 Past Corpus-based Research on the AWL Coverage

The AWL word families have a high frequency of occurrence across multiple written corpora. Coxhead (2000) found that the AWL covers 10% in written academic corpus and 1.4 of the running words in fiction texts. Many other studies have examined the frequency and coverage of the AWL in academic written texts. For example, the AWL covers 11.6% (Cobb & Horst, 2004) and 10.6% (Hyland & Tse, 2007) of different disciplinary corpus (e.g., *linguistics, history, sociology, history*, and *zoology*). This research confirms the supportive role of the AWL across different academic disciplines and supports Nation's (2001) argument that academic vocabulary accounts for 10% of tokens in academic written texts.

It is important to note that the AWL words are not evenly distributed across subject areas. Coxhead (2000) studied the coverage of the AWL in specific disciplines, namely law (9.40%), commerce (12%), art (9.30%), and science (9.10%). In Hyland and Tse's (2007) study, the coverage text of the AWL was 11.10% in engineering, 11% in social sciences, and 9.30% in hard sciences. Chen and Ge (2007) reported that the AWL accounted for 10.07% of their medical corpus containing 190,425 running words. Vongpumivitch, Huang, and Chang (2009), studied

the use of the AWL word families in applied linguistics research articles and found that the AWL provided 11.17% of their 1.5-million token corpus.

There are few studies looking at the distribution of the AWL in spoken academic English. In a corpus-driven study, Hincks (2003) analyzed the nature of the academic vocabulary in Swedish learners' oral presentations and found that the AWL covers only 2.4% of the 13,471 tokens in the corpus. Looking at academic lectures in the BASE corpus, Thompson (2006) found that the AWL provides only 4.9% of running words in the lectures. In a recent study, Paribakht and Webb (2016) conducted a lexical research on the coverage of the AWL word families in listening passages in an English proficiency test used for admission into university. The results revealed that the AWL accounted for only 6.48% coverage of the listening comprehension passages. Similar results were reported by Dang and Webb (2014) who looked at the coverage of Coxhead's (2000) AWL in the BASE corpus. The researchers found that the AWL accounted for only 4.41% of their entire corpus.

Overall, the findings of these studies demonstrate that the coverage figures of the AWL in academic speech are much lower than the figures reported in academic written discourse. This supports the view suggesting that written English differs considerably from spoken English (Halliday, 1979). The next section will shed light on the difference between these two registers.

2.4 Spoken and Written Vocabulary

Given the various communicative purposes, lexical features of spoken and written vocabulary greatly differ (Halliday, 1979). Compared to spoken language, academic written language generally tends to be more formal and contains more complex sentence structures, in that, it includes more lexical diversity, more content vocabulary, and makes use of more discipline-specific vocabulary and low frequency words (Biber, Conrad, Reppen, Byrd, & Helt,

2002; DeVito, 1967; Hasan, 1984; Kroll, 1979; Lee, 2001; Pikulski & Templeton, 2004). This view was supported by Biber's (1986) study who analyzed 41 linguistic features in 545 English text samples of more than 1 million words gathered from different spoken and written texts. Biber (1986) concluded that although written texts make use of more complex structures than spoken texts, this distinction should be viewed as constituting a continuum. In other words, despite the differences between spoken and written texts in terms of topic and some communicative purposes, there is a wide range of overlapping high-frequency vocabulary that is common to both spoken and written discourse. As Biber (1986) notes that "few (if any) absolute differences exist between speech and writing" (p. 385).

Additional evidence on the differences between English spoken and written vocabulary has been provided by corpus-based studies. For instance, Nation (2006) points out that "if 98% coverage of a text is needed for unassisted comprehension, then a 8,000 to 9,000 word-family vocabulary is needed for comprehension of written text and a vocabulary of 6,000 to 7,000 for spoken text" (p.59). Moreover, Dang and Webb (2014) examined the coverage of the AWL in academic spoken corpus (the BASE). Their study shows lower coverage of the AWL in academic spoken English (4.41%) compared to its coverage in academic written corpus (10%), confirming, then, the variation in the vocabulary between written and academic spoken English..

In light of these differences between written and spoken language in terms of the vocabulary aspects, the question arises if any differences can be found within the same register such as academic speech. This study is primarily concerned with academic spoken English, namely lectures and seminars. The next section will present a brief description of the most common aspects of these two speech events.

2.5 Lectures and Seminars

Lectures have some distinctive characteristics making them different from seminars. While lectures by their monologic nature are generally delivered to at least 40 learners in a classroom (Hyland, 2009; Lynch, 2011; Rodgers & Webb, 2016), seminars by their dialogue nature are typically about group discussion and interaction (Carey, 1999; Fielder, 2011; Jordan, 1997), and the classrooms in which seminars take place tend to be much smaller in size. Another distinction between lectures and seminars can be made on the basis of their aims. The primary objective of lectures is to present a wealth of ideas and information through different communicative strategies (e.g., *visuals, gestures*), apart from speech to help learners understand the content of lectures (Bamford, 2004; Basturkmen, 2016; McNeill, 1992; Rowley-Jolivet, 2002; Sharpe, 2006; Sueyoshi & Hardison, 2005). However, given the nature of the various seminar types (e.g., *discussion-based seminars, group-work seminars*) the primary objective of seminars is to actively engage students to enable deep negotiation and discussion of their ideas and information with the aim of optimal understanding (Basturkmen, 1996, 2016; Furneaux, Locke, Robinson, & Tonkyn, 1991).

From pedagogical point of view, the fundamental role of academic lectures and seminars in higher education is well recognized (Kiewra, 2002; Lee, 2009). Both lectures and seminars have been the key methods of university instruction in many countries world-wide (Feak, 2013; Flowerdew, 1994; Long & Richards, 1994; Miller, 2002; Nesi, 2012; Swales, 2001). EAP learners, nevertheless, have difficulties in getting full command of these two academic spoken registers likely due to their inadequate vocabulary knowledge (Basturkmen, 2016; Buck, 2001; Rodgers & Webb, 2016; Stahr, 2009). It has been revealed that there is a strong link between L2 learners' lexical knowledge and their comprehension of written and aural English texts (Bonk

2000; Goh, 2013; Milton, Wade, & Hopkins, 2010; Stahr, 2009; Van Zeeland & Schmitt, 2013; Vidal, 2011). Yet, few attempts have been made to explore EAP learners' vocabulary abilities in relation to the number of words they need to deal with academic English discourse (Dang & Webb, 2014; Webb & Paribakht, 2015) (see section 2.6.3).

This section presented a brief overview of the features of written and spoken registers from the perspective of vocabulary. In the next section the effect of quantity of learners' vocabulary knowledge on their comprehension of different texts will be outlined.

2.6 Vocabulary Knowledge and Comprehension

The following review consists of research that addresses the lexical coverage of various types of English discourse. Lexical coverage is referred to as the percentage of the vocabulary of spoken or written texts known by a learner in order to adequately understand different English texts (Nation, 2006). With the aim of setting particular learning goals, many studies have attempted to determine how much vocabulary learners need to know to understand an English text (see, for example, Laufer, 1989; Schmitt, Jiang, & Grabe, 2011; Webb & Rodgers, 2009a, 2009b). This can be reached by identifying either the lexical coverage necessary for comprehension of English written and spoken texts or the vocabulary size (the number of words) learners should know to be able to understand written and spoken language.

A number of studies have been carried out to determine the lexical coverage an L2 learner needs in order to gain appropriate comprehension of both written and spoken text types (e.g., Laufer, 1989; Nation, 2006). One primary approach that has been used more commonly to determine the adequate threshold for text comprehension is corpus analysis. In the literature related to lexical coverage, the two frequently referenced coverage thresholds are 95% and 98% (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt, 2008; Webb

& Rodgers, 2009). These two estimates are believed to suggest two different degrees of comprehension. While 95% coverage may indicate reasonable comprehension of written texts (Laufer, 1989), 98% coverage may suggest ideal comprehension of written texts (Nation, 2001, 2006). In general, there are two main types of texts that have been examined, written and spoken.

2.6.1 The Relationship of Learners' Vocabulary Knowledge to L2 Comprehension of Written Texts

Much of L2 research examining the coverage needed for comprehension has focused on written discourse (e.g., Hirsh & Nation, 1992; Laufer, 1989; Nation, 2013). Considering the lexical coverage necessary for comprehension of academic texts, Laufer (1989) found that at least 95% coverage is required to reach adequate comprehension of general academic reading texts. The amount of vocabulary needed to reach 95% is approximately 4,000 word families. Coxhead and Nation (2001) indicated how EAP learners can obtain this coverage when they argued that

> Knowing the 2,000 high frequency words and the words in the academic word list will give close to 90% coverage of the running words in most academic texts. When this is supplemented by proper nouns and technical vocabulary, learners will approach the critical 95% coverage threshold needed for reading. (p. 260)

However, Hirsh and Nation (1992) stated that for pleasure reading, learners would need to know around 5,000 word-families to cover 98% of the text. Similarly, Hu and Nation (2000) suggested that 98% coverage is needed for L2 learners to effectively comprehend a fiction text. They stress that one could not have adequate comprehension of a reading passage at only 80% coverage and only some students had an adequate level of comprehension at 90%.

Nation (2006) examined the vocabulary size necessary to reach 95% and 98% coverage. He pointed out that while knowledge of the most frequent 2,000 and 4,000 word families plus proper nouns (mostly geographic and personal names) and marginal words (*uh*, *mmm*) would provide approximately 95% and 98% coverage of a graded reader, the most frequent 4,000 and 8,000 word families were needed to reach 95% and 98% coverage for a newspaper. Schmitt, Jiang, and Grabe (2011) reported a linear relationship between text lexical coverage and learners' comprehension scores. Although the authors found no absolute lexical coverage threshold for comprehension, they indicated that the vocabulary size learners should have varies depending on the degree of comprehension desired. Schmitt et al. (2011) demonstrate that vocabulary coverage of 95% may be necessary if 60% comprehension is targeted; however, 98% coverage is probably necessary if comprehension test scores of higher than 60% is needed. This finding corroborates Laufer and Ravenhorst-Kalovski's (2010) claim who suggested two lexical coverage figures. The first is 95% coverage representing the lower limit needed to gain minimal comprehension, and the second is 98% which represents the upper limit over which learners may gain an optimal level of comprehension.

2.6.2 The Relationship of Learners' Vocabulary Knowledge to Comprehension of L2 General Spoken Texts

While there is a great deal of research that has looked at the lexical coverage of written texts, few studies have targeted the lexical coverage in spoken English discourse. Based on an analysis of a written recall test and a dictation test of four general listening passages, Bonk (2000) found that coverage between 80%–89% may provide adequate listening comprehension. It was concluded that L2 learners using effective coping strategies could achieve adequate L2 listening comprehension of short listening texts at far below 95% coverage. Yet, Schmitt's

(2008) further analysis of Bonk's results revealed that learners with coverage of less than 90% may have had inadequate comprehension while those with knowledge of 95% or more had good comprehension. To examine general spoken discourse, Stæhr (2009) studied the impact of vocabulary size and depth of vocabulary on listening comprehension. His study revealed that a lexical coverage of 98% is needed to effectively deal with spoken language. Stæhr (2009) found that learners who knew the 5,000 word families resulting in 98% lexical coverage had a score of 72.9% in the listening comprehension test, whereas those who mastered the 10,000 vocabulary level providing them with 99.27% lexical coverage of the discourse achieved a score of 80% in the comprehension test.

Van-Zeeland and Schmitt (2012) extended the research examining the connection between lexical coverage and listening comprehension. The researchers studied first language (L1) and L2 learners' comprehension of spoken informal narrative passages. They found that the lexical coverage needed for listening comprehension relies on required degree of comprehension. Van-Zeeland and Schmitt (2012) concluded that while 98% may be "a good coverage target" for ''very high comprehension'' (p. 18), 95% lexical coverage may be adequate for listening comprehension.

Webb and Rodgers (2009a) investigated the vocabulary demand of 318 film scripts of British and American movies, whereas Webb and Rodgers (2009b) examined the lexical demand of 88 British and American TV programs. The researchers found that in order to reach 95% coverage in both the movie and the TV corpora, L2 learners need to know the most frequent 3000 word families. For 98% coverage of English television programs, knowledge of 7000 word families is required and movie corpus needed 6000 word families. Together, these results suggest

that L2 learners would have sufficient lexical knowledge to access conversational English if they know the most frequent 3,000 word families.

In sum, comparing the coverage figures for the written and listening texts shows that the estimates for comprehension written texts are higher than those of general listening texts. This indicates that written language is more demanding for L2 learners, needing exponentially larger vocabulary sizes to reach the two coverage points, 95% and 98%.

2.6.3 The Relationship of Learners' Vocabulary Knowledge to Comprehension of L2 Academic Spoken Texts

As noted above, the research conducted on coverage of English texts has predominately addressed the coverage of written and everyday spoken English rather than academic spoken language. Studies on written and general listening materials may provide some indication regarding the appropriate coverage figures for comprehension of academic spoken English. However, comprehension of academic spoken English may be more challenging than comprehension of general conversation (Van-Zeeland & Schmitt, 2012) due to large amounts of high frequency vocabulary often used in informal conversation compared to those utilized in academic spoken English (Dang &Webb, 2014). Further, academic written English and academic spoken English are dissimilar because they involve different lexical and linguistic features (e.g., *grammatical constructions*). As Swales noted (2001), academic speaking is "much more variable in structure, function, and style than academic writing" (pp. 34–35). This would entail that the threshold needed to comprehend academic writing is different from that of academic speaking.

There is little research that has directly investigated spoken academic language. There are two recent studies that examined the lexical coverage of academic spoken language. Based on 95% and 98% lexical coverage, Dang and Webb (2014) suggest that a vocabulary size of

between 4000-8000 word families plus proper nouns and marginal words is required to comprehend academic spoken English (see section 2.7 for a more detailed description about Dang & Webb's study). When compared to the vocabulary knowledge of 3,000 word families and 6,000-7,000 word families required to reach 95% and 98% coverage, respectively, of general English speech (Nation, 2006; Webb & Rodgers, 2009a, 2009b), Dang and Webb's (2014) findings suggest that L2 learners need a larger vocabulary size to understand academic spoken English than to understand general spoken English. Similar results were reported by Paribakht and Webb (2016) who examined the lexical demand of listening passages in an English proficiency test used for admission into university. They reported that on average, a vocabulary size of 4,000 word families provides 95% text coverage, and a lexical knowledge of 10,000 words would be necessary to gain 98% of the running words in a listening text.

This section dealt with the size of vocabulary learners need to have sufficient comprehension of a text. The findings of the studies mentioned above suggest that the lexical coverage needed for comprehension of written and spoken language vary considerably. The different estimates may arise from the type of input, the nature of the corpus used and the desired degrees of comprehension (Nation, 2001; Nation & Webb, 2011; Schmitt, 2010; Van-Zeeland & Schmitt, 2012; Webb & Rodgers, 2009). Generally, a larger vocabulary size shows better comprehension of a text, and hence greater possible learning gains. It should be noted that the previous studies have also examined the vocabulary size based on the two coverage points, 95% and 98%. These thresholds have been widely used by researchers to represent the minimal and optimal vocabulary size for adequate comprehension of different English texts (see, for example, Dang & Webb, 2014; Laufer, 2010; Nation, 2013; Schmitt, 2010; Webb & Rodgers, 2009). Consequently, this study will use these percentage points to investigate the vocabulary size

necessary for comprehension of academic English. The remainder of this literature review will look at Dang & Webb's study as it is the only study that considered the vocabulary demands of academic spoken English, and therefore it provides a theoretical background for this study.

2.7 Dang & Webb's (2014) Study

Based on the BASE corpus, Dang and Webb (2014) carried out pivotal research to examine the vocabulary size necessary to reach the two threshold points, 95% and 98%, comprehension of academic spoken language. Using the RANGE program and the BNC 14,000 English word lists (Nation, 2006), the researchers analyzed the vocabulary in the BASE corpus consisting of 160 lectures and 39 seminars. They found that while the most frequent 4,000 word families plus proper nouns and marginal words provided 96.05% coverage, knowledge of the most frequent 8,000 word families plus proper nouns and marginal words accounted for 98.00% coverage of general academic spoken English. Dang and Webb also studied the vocabulary size required to gain 95% and 98% coverage of different academic subject areas. They revealed that 5,000 and 13,000 word families give a lexical coverage of 95% and 98% in the life and medical sciences subject. However, vocabulary size of 3,000 word families contribute to 95% lexical coverage and knowledge of 5,000 word families is required to reach 98% coverage of the running words in the social sciences corpus. Given these findings, it seems that specific domains require different vocabulary sizes from general English.

In order to examine the value of the AWL list in academic spoken language, Dang and Webb (2014) further examined the lexical coverage of the AWL in the BASE corpus and across different disciplines. It was found that the AWL coverage ranges from 3.82% of the tokens in the Arts and Humanities sub-corpus to 5.21% in the of Social Sciences sub-corpus, with coverage of 4.41% of the whole BASE corpus. This low coverage of the AWL (4.41%) compared to its high

coverage in written corpus (10%) highlights the lexical variations between spoken and written academic English.

It is worthwhile to note that Dang and Webb's (2014) study was carried out with data from the BASE corpus which is spoken academic British. Therefore, their findings may not be generalised to other regional variations. However, in their recommendations, Dang and Webb (2014) point out that further research is needed on different varieties of spoken academic English to provide a thorough picture on the vocabulary size that may be needed to adequately understand academic spoken texts, and to reach a sufficient ground regarding the lexical coverage of the AWL word families in academic spoken English. Hence, this study looks at the lexical coverage of the AWL in the lectures and seminars presented in the MICASE corpus and the vocabulary demands of American academic spoken English.

2.8 The Present Study

As outlined earlier, based on the two percentage amounts 95% and 98%, a great deal of attention has been paid to determine the vocabulary size needed for comprehension of different text types. However, there is a paucity of research on the vocabulary in academic spoken language (Dang & Webb, 2014). Hence, the present study aims at exploring the vocabulary demands of academic spoken English. There are two major parts to this exploration, the vocabulary profile of academic speech, and lexical coverage of the AWL in academic spoken English. The following primary research questions will drive this study:

1. How many words do learners need to know to reach 95% and 98% coverage of both lectures and seminars presented in academic spoken English (MICASE corpus)?
2. What vocabulary size is necessary to reach 95% and 98% coverage in lectures and seminars separately?

3. What is the coverage of the AWL in both lectures and seminars presented in the academic spoken English?
4. Is the coverage of the AWL in lectures different from that in seminars?
5. With the knowledge of the AWL, what is the vocabulary size needed to reach 95% and 98% coverage of both lectures and seminars presented in the academic spoken English?

2. 9 Methodology

A considerable number of studies have investigated the effective role academic vocabulary plays in preparing EAP learners to achieve great success in university classroom contexts (Ferris, 2009; Gee, 1990; Knodt, 2006; Sullivan, 2006; White, 2007). The previous chapter provided evidence on the importance of academic vocabulary as an indicator of successful L2 comprehension (Gonzalez, 2013; Grobe, 1981; Laufer & Nation, 1995; Nation, 2006; Rodgers & Webb, 2016). The current study uses the corpus-based approach along with certain corpus linguistics software to explore the nature of academic vocabulary used in lectures and seminars of the MICASE corpus. This section presents a detailed description of the research method. This includes information about corpus linguistics as a methodology, the steps taken in data compilation and data selection as well as the research tools and procedures followed for the analysis of the data.

2.9.1 Corpus Linguistics

A corpus (plural 'corpora') is "a collection of texts that has been compiled to represent a particular use of a language" and corpus linguistics "is the compilation and analysis of corpora" (Cheng, 2012, p.6). According to McEnery and Hardie (2011) corpus linguistics "is not a monolithic, [but] consensually agreed set of methods and procedures for the exploration of

language" (p.2). These definitions indicate that corpus linguistics is viewed as a methodology for studying and analyzing the language compiled in a corpus. Yet, the definition of corpus linguistics has been debated. While some researchers argue that corpus linguistics is a discipline, others consider it a methodology (Taylor, 2008; Tognini-Bonelli, 2001).This distinction results in two different approaches used in corpus linguistic research: corpus-based approach and corpus-driven approach (Hunston, Francis, & Manning, 1996; Leech, 1991). Corpus-based approach is used by the researchers who view corpus linguistics as a tool or methodology to "expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study" (Tognini-Bonelli, 2001, p.65), while the researchers who see corpus linguistics as a discipline use the term corpus-driven approach which considers "the corpus itself as a source for hypothesis about language" (Mcenery & Hardie, 2011, p.3). Hence, this study adopts a corpus-based approach because it uses corpus linguistics as a method to derive statistical data related to vocabulary demands of academic spoken English based on a pre-existing academic vocabulary list (the AWL).

Moreover, Cook (2003) states that corpus linguistics research focuses on "the patterns and regularities of language use" (p. 73). In other words, it is concerned with "how speakers and writers exploit the resources of their language" (p. 1) rather than what language use is correct or incorrect (Biber, Conrad, & Reppen, 1998). According to McEnery and Hardie (2011) corpus linguistics refers to the investigations that deal with "some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions" (p.1).

Corpora can contain a wide variety of academic written and/or spoken language texts and come from different resources, for example, academic articles and lectures. Two key features of the language included in corpora is that it is electronically stored and occurs naturally as

indicated by Conrad (2002) “a corpus is a large, principled collection of naturally occurring texts that is stored in electronic form (accessible on computer)” (p. 76). Considering these descriptions, it can be assumed that a corpus is an electronic collection of naturally occurring language samples compiled to investigate a particular use of language. For the present study, the authentic language samples are academic lectures and seminars of the MICASE corpus compiled electronically from university classrooms.

One key aspect of a corpus is that it can be analyzed by the use of corpus linguistics computer programs as Cheng (2012) noted that “[a corpus] is made accessible by means of corpus linguistic software that allows the user to search for a variety of language features” (p. 6). Flowerdew (2004) highlights the significant role of corpus analysis in providing “attested examples of recurring language patterns which are based on empirical data” (pp.13-14). Many researchers agree that corpus linguistics studies can offer empirical analysis of language features across different text types of corpora, and therefore provide better understanding of language patterns and use (Connor & Upton, 2004; O'Keeffe & McCarthy, 2010; Stubbs, 2007). This study used an analysis software tool to examine and identify the nature of vocabulary in academic spoken corpus.

With respect to the type of research method conducted for the present study, it can be categorized as quantitative for two main reasons. First, the numeric data collected by the computer software were based on the AWL word list and the MICASE corpus. Second, the empirical data produced were used to assess the coverage percentages of the AWL in the target corpus. However, unlike other quantitative studies that were designed to test pre-existing hypotheses, this study is exploratory in nature and has a non-experimental design as the data collection was based on naturally occurring speech and basically concerned with *what* the lexical

demand of academic speech rather than *why* or *how* academic vocabulary is used by those speakers of academic English.

Overall, corpus linguistics research relies on computer software to analyze computerized corpora in order to develop the knowledge and explanation pertaining to different language patterns. In the case of this study, AntWordProfiler program (Anthony, 2014) was used to examine the usage patterns of the AWL words across the lectures and seminars represented in the MICASE corpus.

Having looked at corpus linguistics as a methodology and the rationale for using a corpus-based approach in the current study, the following section provides reasons for using a specialized corpus.

2.9.2 Categories of Corpora

There are two broad types of corpora: general and specialized. A general corpus tends to be large and aims to represent the language variation of the whole speech community (Cheng, 2012; Flowerdew, 2004). For instance, the 100 million-word BNC consists of a wide variety of spoken and written texts which is believed to represent the entire variety of British English (Leech, Rayson, & Wilson, 2001). A specialized corpus, (e.g., the MICASE corpus) on the other hand, is smaller in size and can be used "as a means of discovering the characteristics of a particular area of language use" (Aston, 1997, p. 61).

It has been assumed that the size of the corpus relies on the purpose it is used for (Flowerdew, 2004; Sinclair, 1991). Because a small specialized corpus is often derived from a particular type of texts, it generally provides a better understanding of the linguistic feature being studied (Connor & Upton, 2004). Furthermore, compared with a general corpus, a smaller specialized corpus can be more useful for English for Specific Purposes (ESP) and EAP learners

as it is more relevant to the contexts where corpus texts were produced (Flowerdew, 2009; Tribble, 2002). Furthermore, Hunston (2002) argues that a smaller specified corpus aims "to be representative of a given type of text and it is used to investigate a particular type of language" (p.14). However, for the purpose of this study, more than one kind of text was investigated, but the texts used belong to the same language register, that is, spoken academic English. Finally, Flowerdew (2004) argues that a specialized corpus can be more appropriate than a general corpus for understanding a particular language of a specific academic nature. Given the above arguments, it can be assumed that the MICASE corpus would be the best choice to explore the coverage of academic vocabulary in lectures and seminars because the MICASE is, in fact, a specialized corpus in the sense that it contains only academic spoken language and can be studied on the basis of more specific genres (text types) such as lectures or seminars.

The previous section outlined the case for using a specialized corpus (MICASE) in this study. An overview of the characteristics of this corpus will be presented in the following section.

2.9.3 Selection of Materials

This section provides information pertaining to the materials selection, including information about the corpus, features of the corpus samples and the wordlist employed in this study.

2.9.3.1 Corpus Selection

The corpus used for this study was the MICASE, an academic spoken language corpus of roughly 1.8-million words (approximately 200 hours) of contemporary speech recorded at the University of Michigan between 1997-2002 (see Simpson, Briggs, Ovens, & Swales, 1999). It covers a broad range of speech events (totaling 152 speech events) classified into classroom

events (such as lectures, seminars, and discussion sessions) and non-classroom events (such as tours and tutorials, study groups, interviews and office hours), with number of words ranging from 2,805 to 30,325 (Simpson-Vlach & Leicher, 2006). The speakers of the corpus represent almost the whole university, including faculty, staff and students at all levels representing native speakers (88%) and non-native speakers (12%). Given the assumption that language patterns in an academic setting differ from those found in informal spoken or academic written discourse, the English Language Institute at the University of Michigan envisioned the MICASE corpus as a major source that can be used for research purposes and for the development of English teaching materials of EAP programs and tests. The MICASE corpus is freely available at http://micase.umdl.umich.edu/m/micase/.

With respect to the corpus samples, this study was based on specific text genres (types), namely lectures and seminars included in the MICASE corpus. These academic speech genres were selected for three primary reasons. First, lectures and seminars still remain the overriding method of teaching in English-medium universities worldwide (Hyland, 2013; Jones, 2007; Rodgers & Webb, 2016) and have so far been a fruitful area of investigation for a substantial number of researchers (e.g., Biber, 2009; Furneaux, Locke, Robinson, & Tonkyn, 1991; Hyland, 2006; Morita, 2004; Nesi & Basturkmen, 2006; Simpson-Vlach & Ellis, 2010; Swales, 2001). Second, although there are several studies investigating specifically the MICASE lectures and seminars (e.g., Basturkmen, 2016*;* Crawford Camiciottoli, 2004; Fortanet, 2004; Lee, 2006; Okamura 2009; Skyrme, 2010; Thompson, 1994), none of these studies examined the lexical profile of these two university genres and thus, this study was the first study examining the lexical demands of American spoken academic English. It should be noted that Dang and Webb (2014) studied the lexical demands of lectures and seminars of British academic English.

However, this study is based on the analysis of a different academic context, that is, American English, and therefore it fills a gap in the literature. Additionally, because the BASE and MICASE corpora represent two different academic cultures (Lin, 2012; Nesi, 2012; Thompson, 2006), one might expect, for example, discrepancies between these two corpora in terms of the lexical items used. Third, since these university lectures and seminars were prepared in real academic classroom settings, they are more likely to be representative of the academic language EAP learners may encounter at universities where English is introduced as medium of instruction (Dang & Webb, 2014).

2.9.3.2 Text Selection

The MICASE corpus consists of transcripts taken from four academic divisions 1) Social Sciences and Education, 2) Physical Sciences and Engineering, 3) Humanities and Arts, and 4) Biological and Health Sciences (Simpson-Vlach & Leicher, 2006). It is worth noting that not all the transcripts of the MICASE lectures and seminars, which are the focus of the present study, were found in the four divisions. While the 62 lectures transcripts were compiled from the above four divisions, the 7 seminars transcripts were compiled from only two divisions: Social Sciences and Education, and Humanities and Arts. With reference to the MICASE collections, a corpus comprising sixty-two lectures and seven seminars in the four academic divisions was selected. An overview of the source composition of the downloaded transcripts used in this study is provided in Table 1. As shown, the Table demonstrates the number and the type of speech events (seminars and lectures) as well as the number of words in each speech event, which appeared across the four academic divisions.

Table 1

Academic Divisions with Speech Events Types and Numbers Represented in the Current Study (Simpson-Vlach & Leicher, 2006)

	Social Sciences & Education Division			**Biological & Health Sciences**	
	Speech event type	Tokens		Speech event type	Tokens
1	Medical Anthropology Lecture	11,941	1	Intro Biology First Day Lecture	6995
2	Intro Psychology Lecture	7845	2	Drugs of Abuse Lecture	6178
3	Graduate Macroeconomics Lecture	8736	3	Intro to Biochemistry Lecture	11,788
4	Principles in Sociology Lecture	12,371	4	Biology of Cancer Lecture	11,647
5	Behaviour Theory Management Lecture	14,385	5	Race and Human Evolution Lecture	11,366
6	Intro Communication Lecture	9805	6	General Ecology Lecture	6932
7	Media Impact Communication Lecture	9900	7	Intro to Evolution Lecture	12,427
8	Intro Anthropology Lecture	11,654	8	Biology and Ecology of Fishes Lecture	9719
9	Political Science Lecture	15,359	9	Biology of Birds Lecture	12,253
10	Intro to Psychopathology Lecture	8375	10	Biology of Fishes Group Activity	2,866
11	Honors Intro Psychology Lecture	5843	11	Graduate Population Ecology Lecture	5369
12	Sex, Gender and the Body lecture	14,629	12	Graduate Cellular Biotechnology Lecture	13,409
13	Labor Economics Lecture	12,560	13	Microbial Genetics Lecture	13,994
14	Ethics Issue in Journalism Lecture	16,291	14	Spring Ecosystems Lecture	11,651
15	Archaeology of Modern American life lecture	10,924		**Physical Sciences & Engineering**	
16	Statistics in Social Science Lecture	16,748		Speech event type	Tokens
17	Graduate Public Policy Seminar	25,414	1	Intro Engineering Lecture	6651
18	Politics of Higher Education Seminars	19,687	2	Intro Oceanography Lecture	8600
	Arts & Humanities		3	Intro to Physics Lecture	7880
	Speech event type	Tokens	4	Inorganic Chemistry Lecture	6918
1	Literature and Social Change Lecture	10,207	5	Structure and Reactivity II Lecture	4622
2	Japanese Literature Lecture	8676	6	Separation Processes	5438
3	Fantasy in Literature Lecture	13,545	7	Graduate Physics Lecture	14,611
4	Perspectives on the Holocaust Lecture	9258	8	Number Theory Math Lecture	4144
5	History of the American Family Lecture	11,102	9	Professional Mechanical Engineering Seminar	13,180
6	Renaissance to Modern Art History Lecture	8332	10	Graduate Industrial Operations Engineering Lecture	11,098

(Continued)

(Table 1, continued)

	Arts & Humanities			**Physical Sciences & Engineering**	
7	Twentieth Century Arts Lecture	6246	11	Radiological Health Engineering Lecture	13,658
8	Sports and Daily Life in Ancient Rome Lecture	12,958	12	Intro Programming Lecture	8094
9	Historical Linguistics Lecture	12,841	13	Dynamic Earth Lecture	7011
10	Intro. Latin Lecture	5883	14	Rehabilitation Engineering and Technology	7374
11	Graduate Online Search and Database Lecture	20,012	15	Intro. to Groundwater Hydrology Lecture	14,151
12	Women in the Bible Lecture	10,387			
13	Visual Sources Lecture	12,526			
14	American Literature Lecture	16,104			
15	African History Lecture	9290			
16	Beethoven Lecture	7821			
17	Graduate Philosophy Seminar	22,214			
18	Graduate Buddhist Studies Seminar	26,075			
19	Graduate French Cinema Seminar	24,458			
20	First Year Philosophy Seminar	13,906			
21	English Composition Seminar	21,442			

Like other classroom speech events, the MICASE lectures and seminars differ in terms of students' number and interactivity level (Simpson et.al, 1999). The lectures are categorized into small and large depending on "an arbitrary cut-off point (40 students)" (Simpson-Vlach & Leicher, 2006, p. 18). In other words, small lectures consist of 40 students or fewer while the large lectures involve more than 40 students, with some around 400 students. Unlike large lectures which are basically monologic, small lectures can be monologic, interactive or mixed. Regarding seminars, these are generally highly interactive consisting basically of upper-level undergraduate or graduate level students (Simpson-Vlach & Leicher, 2006).

The 62 transcripts in this study also vary from the point of view of contextual information, for example, academic discipline, number and level of participants, duration of the transcript, among other information. These transcripts contain varying topics across the four divisions, such as Macroeconomics, Statistics, and Philosophy. The number of participants in each transcript also varied from 3 to 400 and their level was classified into graduate, undergraduate and mixed faculty, staff and students. For the length of recordings, the MICASE lectures and seminars ranged in length from 36 to 169 minutes. For further information about the characteristics of the lectures and seminars and other MICASE speech events see Simpson-Vlach and Leicher (2006).

It is worthwhile to have these different characteristics related to a wide range of transcripts in order to yield more valid results of the vocabulary coverage. In other words, because this study comprises two different text types (lectures and seminars) covering a considerable number of academic disciplines (69 topics), they provide a good representation of academic lectures and seminars, and hence one can expect more valid results with respect to the coverage of the AWL word families across these disciplines.

2.9.3.3 The Wordlists

The wordlists used in the study were the GSL, the AWL and the BNC/COCA. The GSL was developed by West (1953) and includes a list of high frequency words containing about 2000 basewords. Coxhead (2000) compiled the AWL on top of the GSL which has been criticised for being dated (based on texts published before 1930's). The assumption that the GSL represents general high-frequency English words indicates that the AWL comprises general high-frequency words that are not included in the GSL (Gardner & Davies, 2013). Thus, to see the coverage of the GSL, its first and second 1,000 words were examined in the MICASE lectures and seminars corpus. Besides, the coverage of the GSL obtained in this study corpus was compared with that of Dang and Webb (2014) who examined its distribution in the BASE corpus which is spoken academic British.

With respect to the AWL, it contains 570 word families comprising about 3,082 family members. The AWL was investigated in relation to its lexical coverage in the target corpus and its usefulness for learners with different vocabulary sizes. The AWL is available online from the Victoria University of Wellington, New Zealand at http://www.victoria.ac.nz/lals/resources/academicwordlist/awl-headwords.

With regard to Nation's (2012) BNC/COCA word family lists, these comprise 25,000 word families organised into 1,000-word-family lists plus the four additional lists of proper nouns, transparent compounds, marginal words and abbreviations (see section 2.11.2 for a description of BNC/COCA). The BNC/COCA word family lists are downloadable from Paul Nation's Victoria University Profile https://www.victoria.ac.nz/lals/about/staff/paul-nation.

2.10 Corpus Compilation

This section describes the steps taken to compile the corpus (transcripts) for analysis. It demonstrates the process of preparing the transcripts and the decisions made regarding cleaning the data of the corpus.

2.10.1 Cleaning the Data

Even though all the MICASE texts composing the corpus have been originally stored in electronic format, these texts were in XML files, and hence they are not appropriate for use by corpus analysis software used in this study which can only read TXT files. Therefore, these files had to go through format conversion and cleaning processes to be compatible with the corpus software.

For the present study, the data required a proper treatment before using the AntWordProfiler software. The original HTML files were first downloaded. The scripts were then extracted from their original HTML documents and saved as separate Word (.doc) files. Next, these files were cleaned up by removing extraneous words. For example, the transcripts of the MICASE lectures and seminars corpus contain information about these transcripts found in their headers. Examples of such information are the subject area of the speech event, the type of speech event, as well as the number and level of participants. This type of information was removed from the texts. Extra-linguistic information and non-verbal features such as *<PAUSE>, <SOUND EFFECT>, < LAUGH>, <APPLAUSE>, <BACKGROUND NOISE>, <WHISPERING>* etc. were also excluded from the corpus to ensure reliable analysis. Similarly, the incomplete words (e.g., *th-, cl-, wa-)* were also taken out from the corpus.

Moreover, contractions (e.g., *i'd,* wouldn've) and reduced forms (e.g., *hafta*, *kinda*, *cuz*), as well as the misspellings (*archeologica, nber*) were changed to fit the spelling found in the BNC/COCA wordlists. For example, the words *wouldn've*, *i'd*, *hafta* and *cuz* were changed to *would not have*, *I would*, *have t*o and *because*, respectively. Were the spellings of these words not been changed, there would have been identified as less frequent than Nation's (2012) 25000 most frequent word families. Webb and Rodgers (2009a, 2009b) suggest that contractions and reduced forms can play a role in students' vocabulary learning

and comprehension of spoken texts. The researchers suggest that listeners' knowledge of the expanded forms of the words does not necessarily mean that they understand the contracted forms of the same words. For instance, listeners may recognise *give me* and *will not* but might not know *gimme* and *won't*. However, due to the expected small percentage of contractions, their effect on learners' comprehension would be only minor.

Likewise, many of the hyphenated words were changed into single words after deleting the hyphen. For example, the hyphen was removed from the words *short-term* which were rewritten as two separate words *short term.* In contrast, in acronyms such as C-I-A and F-B-I the hyphen and the spaces were removed and the acronyms spelled out as CIA and FBI. The reason for separating the hyphenated words is because the software can recognize the different word parts and organize them according to BNC/COCA 25,000 frequency levels they refer to, and hence offer a more accurate assessment of the academic vocabulary used in the corpus. For the sake of clarity, an example of a text cleaning of a sample lecture transcript is displayed in Table 2.

Table 2

Examples of a Text of a Transcript already Cleaned for the Study

An unclean text	**A cleaned text**
S1: hang on a second. um, so the point is that scan-F returns a value that evaluates to zero cuz it didn't read anything. one is not equal to zero. that's a true statement so, i get out. something bad happened. <*PAUSE:06*> you nee- you may need to go and actually play with this code and look at it a little to see this. <*STUDENTS GETTING UP TO LEAVE*>	Hang on a second. um, so the point is that scan F returns a value that evaluates to zero because it did not read anything. One is not equal to zero. That is a true statement so, I get out. Something bad happened. You may need to go and actually play with this code and look at it a little to see this.

Note. Available from https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;cc=micase;view=transcript;id=COL485MX069

Finally, after cleansing the corpus, the Word files and the documents were saved as TXT files. Given that one of the questions of this study is concerned with comparing the coverage of the AWL between individual sub-corpora (lectures and seminars), all the texts were submitted to the same file conversion and cleaning processes before being labelled and saved in "plain text", that is TXT format. The (LECTURES.txt) includes all the 62 transcripts of the MICASE lectures, the (SEMINARS. txt) includes all the 7 transcripts of the MICASE seminars and (LECTURES & SEMINARS. txt) is a combination of all the 69 transcripts of the MICASE lectures and seminars. Generally, the whole cleaning process ensures that the corpus can be properly used by the corpus software in order to make the corpus analysis more reliable. Following the data cleaning, the final product of this corpus comprised two sub-corpora (lectures and seminars) with 69 texts, totalling 740,651 running words. Table 3 shows the number of texts and tokens per sub-corpus analyzed in this study.

Table 3

Number of Texts and Words by Sub-Corpus

	Sub-corpus	Texts	Tokens
1	Lectures	62	594,117
2	Seminars	7	146,534
3	Lectures and seminars	69	740,651

2.11 Data Analysis

This section provides a brief description of the research tools used in the analysis of the data and procedure followed for data analysis.

2.11.1 AntWordProfiler

AntWordProfiler, which has supplanted Nation and Heatley's (2002) RANGE program (http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx) is a freeware vocabulary

profiler developed by Anthony (2009) to carry out corpus linguistics research on word profiling. By default, AntWordProfiler (available at www.antlab.sci.waseda.ac.jp) is preloaded with the first and second thousand level wordlists of the GSL, as well as the 570 AWL word families. As such, users can upload their texts of a corpus or sup-corpus to the AntWordProfiler program and compare it with these preselected vocabulary lists to find the word family levels of the texts based on their frequency in the lists, and to find the percentage of the words in the texts that are covered by these established word lists, that is, lexical coverage. For this end, the AntWordProfiler 1.4.0 (Anthony, 2014) along with the BNC/COCA wordlists (Nation, 2012) were used to assess the overall academic vocabulary load of the MICASE lectures and seminars corpus. Further information about the AntWordProfiler software and its analytical tools can be found in the Readme file on Anthony's homepage (readme_file_AntWordProfiler130.txt).

2.11.2 BNC/COCA Lists

The BNC/COCA lists consist of 25th 1000 word-family level grouped in descending frequency order, and four lists of proper nouns (list 26), marginal words (list 27), transparent compounds (e.g., *birthday*) (list 28), and abbreviations (list 29). The BNC/COCA frequency word lists are comprehensive in the sense that they are representative of both British (BNC corpus) and North American (COCA corpus) varieties of English and compiled based on written and spoken corpora.

The 29 BNC/COCA word family lists account for roughly 99% coverage of the running words in different kinds of corpora (Nation & Anthony, 2013), suggesting that these lists include high, mid and low frequency vocabulary. The BNC/COCA lists can be used along with corpus software to provide a quantitative analysis of a corpus in relation to the number of words lists such as the GSL has in that corpus (Nation & Webb, 2011). In this study, the 29 BNC/COCA word family lists were utilized together with AntWordProfiler to

analyze the academic spoken English and to display how many words of the AWL occurred at each 1,000 word frequency level.

2.12 Procedure

In order to identify the vocabulary size level required to reach 95% and 98% coverage, the combined corpus (lectures and seminars as a whole) was run through the AntWordProfiler program against the BNC/COCA word lists. The AntWordProfiler produced the number and percentage of words in the corpus that occur in each 1,000-word frequency level up to 25,000 word family level as well as the number and percentage of the words that appear in the additional four lists (proper nouns, marginal words, transparent compounds, and abbreviations) and the "Not in the list" section. To see if the lexical coverage of the lectures differs from that of seminars, these two sub-corpora were run separately through the AntWordProfiler program against the BNC/COCA word lists and the data obtained were explored and displayed individually. After analyzing the corpus as a whole and separately using the AntWordProfiler and the BNC/COCA word lists, the data were then sorted in Excel worksheets for analysis.

To obtain the coverage of the AWL and the GSL vocabulary lists in the lectures and seminars corpus, this corpus was uploaded as one TXT file and then separately onto the AntWordProfiler program, with the AWL and the GSL serving as the baselists in the program. The AntWordProfiler would give the coverage percentage and the number of the AWL and the GSL words in the uploaded corpus.

To find the distribution of the AWL at different frequency levels, the AWL was run through AntWordProfiler as the text and BNC/COCA lists as the baseword lists. The AntWordProfiler would classify the family words of the AWL into one of the twenty-five 1000-word BNC/COCA frequency levels and into the additional four lists (proper nouns, marginal words, transparent compounds, and abbreviations). The output also shows what

words in the AWL are not in the BNC/COCA lists and hence fall in the "Not in the List" category of the AntWordProfiler. This category indicates very low frequency words, and thus they do not occur in any of the BNC/COCA lists which were used to analyze the AWL word families. It should be noted that although the proper nouns list involved in the BNC/COCA lists includes over 22,400 entries, this number is unlikely to account for all the proper nouns included in a corpus. Hence, the AntWordProfiler organized a number of proper nouns as "Not in the lists". Those proper nouns found in this list were further manually sifted and added to totals of the proper nouns.

To determine the coverage provided by the AWL at each 1,000 word frequency levels, the data had to be analyzed in several steps. First, the AWL was run through the AntWordProfiler as the text with BNC/COCA as the level lists. The AntWordProfiler output distributed the AWL word families at various 1,000-word frequency levels, indicating the lexical profile statistics of the AWL. The AntWordProfiler also displayed the AWL words that are distributed in the other four additional lists and in the "Not in the List" category. Second, the AWL word families appeared at each of the frequency levels were reorganized in a separate sub-list. Finally, the lectures and seminars corpus was uploaded onto the AntWordProfiler as a text with each frequency sub-list generated above as the level lists. The AntWordProfiler output would display the lexical coverage of the AWL sub-lists that occur at various frequency levels in the lectures and seminars corpus. For example, there were 19 AWL words appeared at the first 1,000 word frequency level. These 19 AWL words were reorganised into a sub-list for the first 1,000 word frequency level and served as the level list in the AntWordProfiler. The AntWordProfiler showed that this sub-list accounted for coverage of 0.0942% in the lectures and seminars corpus.

2.13 Findings

This section presents the findings of the corpus analyses carried out in the previous section. Given the research questions (see section 2.8) guiding this study, this section is divided into three main parts. The first depicts the lexical coverage needed for comprehension of the MICASE lectures and seminars. The next part presents the findings related to the coverage of the AWL in academic spoken corpus represented by the MICASE lectures and seminars. The third part of this section assesses the value of the AWL.

2.13.1 Computing the Lexical Coverage of the MICASE Lectures and Seminars

Table 4 displays the cumulative coverage including proper nouns and marginal words for the seminars and lectures corpora, both separately and in combination. Previous studies investigating vocabulary demand of spoken English discourse suggested that proper nouns and marginal words have a minimal learning burden for language learners (Nation, 2006; Webb & Rodgers, 2009a, 2009b), and thus, added the coverage percentage of these words to the potential coverage. As shown, a vocabulary of 3,000 word families plus proper nouns and marginal words would provide 95.55% coverage and vocabulary of 7,000 word families plus proper nouns and marginal words accounted for 98.03% coverage of the combined seminars and lectures corpus. The vocabulary knowledge needed to reach 95% coverage is somewhat similar for both sub-corpora. Including proper nouns and marginal words, knowledge of the most frequent 3,000 word families was needed to reach 95.3% and 96.64% coverage of the lectures sub-corpus and seminars sub-corpus, respectively. However, larger differences were found between the two genres in terms of the vocabulary needed to reach 98% coverage. The vocabulary necessary to reach 98% coverage ranged from 6,000 to 8,000 plus proper nouns and marginal words. With knowledge of the most frequent 6,000 word families plus proper nouns and marginal words, learners would reach 98.29% coverage of the seminars sub-corpus, whereas they may need the most frequent 8,000 word families including proper nouns

and marginal words to reach 98.15% coverage of lectures sub-corpus. The results suggest that at 98% coverage, lectures are more lexically demanding, needing larger vocabulary sizes to reach this coverage point.

Table 4

Cumulative Coverage Plus Proper Nouns and Marginal Words for the MICASE Lectures and Seminars Corpus, both Independently and in Combination

Word list	Lectures and seminars	Lectures	Seminars
1,000	86.44	85.86	88.85
2,000	91.99	91.59	93.7
3,000	95.55[a]	95.3[a]	96.64[a]
4,000	96.67	96.49	97.43
5,000	97.28	97.12	97.96
6,000	97.69	97.56	98.29[b]
7,000	98.03[b]	97.93	98.54
8,000	98.24	98.15[b]	98.71
9,000	98.38	98.29	98.83
10,000	98.55	98.48	98.92
11,000	98.69	98.63	99.02
12,000	98.75	98.7	99.05
13,000	98.82	98.78	99.08
14,000	98.88	98.85	99.11
15,000	98.92	98.89	99.12
16,000	98.97	98.94	99.14
17,000	99.02	99	99.17
18,000	99.07	99.04	99.27
19,000	99.09	99.06	99.28
20,000	99.12	99.09	99.29
21,000	99.13	99.1	99.29
22,000	99.14	99.11	99.3
23,000	99.16	99.13	99.31
24,000	99.16	99.13	99.31
25,000	99.17	99.14	99.31
Proper Nouns	1.21	1.25	1.09
Marginal Words	1.75	1.76	1.72
Abbreviations	0.16	0.17	0.16
Transparent Compounds	0.18	0.19	0.14
Not in the list	0.47	0.49	0.4
Total	740,651	594,117	146,534

[a] Reaching 95% coverage
[b] Reaching 98% coverage

2.13.2 The AWL Coverage

The coverage of the AWL in the lectures and seminars corpus, both separately and in combination is shown in Table 5. The AWL accounted for 3.68% coverage of the combination of lectures and seminars included in the MICASE corpus. The list was not evenly distributed between the two sub-corpora. The AWL had higher coverage in lectures sub-corpus (3.80%) than in seminars sub-corpus (3.15%).

Table 5

Coverage of the MICASE Lectures and Seminars Corpus, both Separately and in Combination, by the GSL (West, 1953) and AWL (Coxhead, 2000) (%.)

Corpus	Proper nouns	Marginal words	General Service List		AWL
			1st 1,000 words	2nd 1,000 words	
Lectures	1.25	1.76	82.99	3.61	3.80
Seminars	1.09	1.72	85.44	3.46	3.15
Lectures and seminars	1.21	1.75	83.47	3.58	3.68

It has been found that some AWL word families appear at the 1st 1,000, 2nd 1,000 and 3rd 1,000 word-frequency levels of some word family lists (Cobb, 2010; Dang & Webb, 2014; Nation, 2004). As a result, it is essential to investigate the lexical size of the AWL in the BNC/COCA word lists in order to determine the vocabulary size necessary to reach 95% and 98% coverage of the academic spoken English with the support of the AWL. Table 6 presents the lexical profile statistics of the AWL word families across the 1st-25th 1,000 word-family lists. It shows that 19 AWL word families (1,252 tokens) occurred in the first BNC/COCA 1,000-word list. This provided 0.0942% coverage of the MICASE lectures and seminars corpus. 134 AWL word families (5,728 tokens) appeared in the second 1,000 BNC/COCA word list, and accounted for 0.6641% coverage of the MICASE corpus. 327 AWL word families (7,223 tokens) were in the third BNC/COCA 1,000-word list, providing 1.6207% coverage of the MICASE corpus. The number of AWL items in the fourth, fifth, sixth and seventh BNC/COCA 1,000-word lists was 70 (555 tokens), 27 (166 tokens), 10 (56

tokens) and 5 (57 tokens). These items accounted for coverage of 0.3469%, 0.1338%, 0.0496 and 0.0248, respectively. From the eighth to the eleventh 1,000 word level, very few AWL word families occurred (1 (2 tokens), 2 (5 tokens),3 (3 tokens), 1 (3 tokens), respectively), which provided the corresponding coverage in the MICASE corpus: 0.005 %, 0.0099 %, 0.0149 %, 0.005 %. From twelfth to twenty-fifth 1,000-word level, none of the AWL word families appeared.

In addition to the distribution of AWL word families across the twenty-five BNC/COCA lists, one word family (2 token) was in the BNC/COCA wordlists of transparent compounds (*widespread*), accounting for 0.005% coverage of the MICASE lectures and seminars corpus.

Table 6

Lexical Profile Statistics of the AWL

Frequency levels[1]	Raw Tokens	Word family	
		No	Percentage (%)
1,000	1,252	19	0.0942
2,000	5,728	134	0.6641
3,000	7,223	327	1.6207
4,000	555	70	0.3469
5,000	166	27	0.1338
6,000	56	10	0.0496
7,000	57	5	0.0248
8,000	2	1	0.005
9,000	5	2	0.0099
10,000	3	3	0.0149
11,000	3	1	0.005
Transparent compound	2	1	0.005
Total	725,599	20,177	2.9739

Note: Frequency levels refer to word families of the BNC/COCA word lists.

2.13.3 The Value of the AWL

Table 7 shows the support provided by the AWL for learners with different amounts of vocabulary as defined by the 25 BNC/COCA 1,000 word family lists. The second column of Table 7 demonstrates the cumulative coverage of the AWL words in the MICASE lectures

and seminars corpus at each level of the BNC/COCA list. This cumulative coverage is calculated by adding each coverage point of the AWL items to the sum of its predecessors. For example, the cumulative coverage of the AWL for the 2nd 1,000 word level was 0.7583% (0.0942% + 0.6641 % = 0.7583%), and the cumulative coverage of the AWL for the 3rd 1,000 word level was 2.379% (0.0942% + 0.6641% + 1.6207% = 2.379%). Cumulative coverage of the AWL indicates the proportion of the AWL items that are included in the most frequent 1,000 to 25,000 BNC/COCA word families. The third column of Table 7 points to the supportive coverage the AWL items provided to learners who are at the lower frequency level. The supportive coverage was calculated by subtracting the cumulative coverage provided by the AWL items at each BNC/COCA word level from the total coverage of the AWL in the MICASE corpus (3.6848). For instance, assuming that learners knew the most frequent 1,000 BNC/COCA word families together with the AWL, their knowledge of the AWL would account for 3.5906% coverage in academic spoken texts (3.6848-0.0942%=3.5906%). Likewise, for learners with the knowledge of the second, third, fourth, fifth and sixth 1,000 BNC/COCA word families, the corresponding coverage that the AWL provided would be 2.9265%, 1.3058%, 0.9589%, 0.8251% and 0.7755%, respectively. From the seventh to eleventh BNC/COCA word level the AWL had average supportive coverage of 0.7338%.

Table 7

Supportive Coverage Provided by the AWL List for Learners with Different Lexical Sizes as Determined by BNC/COCA Word Lists (%)

Word list	Cumulative coverage of the AWL items in the BNC/COCA word lists	Supportive coverage of the AWL
1,000	0.0942	3.5906
2,000	0.7583	2.9265
3,000	2.379	1.3058
4,000	2.7259	0.9589
5,000	2.8597	0.8251
6,000	2.9093	0.7755
7,000	2.9341	0.7507
8,000	2.9391	0.7457
9,000	2.949	0.7358
10,000	2.9639	0.7209
11,000	2.9689	0.7159
Transparent Compounds	2.9739	0.7109

Table 8 shows the potential coverage that learners with different lexical sizes may achieve with the help of the AWL as defined by BNC/COCA. The third column of Table 8 demonstrates the potential coverage that learners may reach by learning the AWL items. The potential coverage for these learners at a given BNC/COCA word level was the sum of the cumulative coverage including proper nouns and marginal words at that word level and the supportive coverage provided by the AWL items at the next lower frequency level. For example, if learners know the most frequent 1,000 BNC/COCA word families and study the AWL, they may reach potential coverage of 90.03% (86.44% + 3.5906%). Learners with the vocabulary level of the most frequent 3,000 word families in the BNC/COCA may reach 96.86% coverage with the help of the AWL. As shown in Table 8, learners who have

mastered the most frequent 5,000 BNC/COCA word families and the AWL will have 98.11% coverage of academic spoken English. More importantly, for learners with a vocabulary size of the most frequent 5,000 BNC/COCA word families, knowledge of the AWL may enable them to reach optimal comprehension (98.11%) of academic spoken lectures and seminars.

Table 8

Potential Coverage Learners of Different Lexical Sizes may Reach with Knowledge of the AWL as Defined by BNC/COCA Word Lists

Word list	Cumulative coverage including proper nouns and marginal words for the MICASE lectures and seminars without knowledge of the AWL	Potential coverage with knowledge of the AWL
1,000	86.44	90.03
2,000	91.99	94.92
3,000	95.55[a]	96.86[a]
4,000	96.67	97.63
5,000	97.28	98.11[b]
6,000	97.69	98.47
7,000	98.03[b]	98.78
8,000	98.24	98.99
9,000	98.38	99.12
10,000	98.55	99.27
11,000	98.69	99.41
Proper Nouns	1.21	
Marginal Words	1.75	
Abbreviations	0.16	
Transparent Compounds	0.18	
Not in the list	0.47	
Tokens	740,651	

[a] Reaching 95% coverage.
[b] Reaching 98% coverage.

2.14 Discussion

This section discusses the main findings in light of the questions stated in section 2.8. These findings will be compared with prior studies and relevant literature where appropriate.

2.14 .1 Vocabulary Size

This section aims to discuss the first two research questions of this study.

1. **How many words do learners need to know to reach 95% and 98% coverage of both lectures and seminars presented in the academic spoken English?**
2. **Is there a difference between the vocabulary size necessary to reach 95% and 98% coverage of lectures and seminars?**

In response to the first question, the findings (as presented in Table 4) revealed that a vocabulary size of the most frequent 3,000 word families plus proper nouns and marginal words would provide 95.55% coverage, and a vocabulary size of the most frequent 7,000 word families plus proper nouns and marginal words would provide 98.03% coverage of American academic lectures and seminars corpus.

The coverage figures in the present study turn out to be different from the findings of Dang and Webb (2014), who concluded that it was necessary to know 4,000 word families to reach 95% coverage and 8,000 word families to reach 98% coverage of lectures and seminars of British Academic spoken English. This suggests that American lectures and seminars may be less lexically demanding for learners than British lectures and seminars. There may be a number of reasons for the difference between the present study and Dang and Webb's (2014) with respect to the vocabulary sizes needed to achieve these figures. The first reason is likely related to the nature of university lectures and seminars in both corpora. Practices of American university classrooms, especially lectures show higher level of interaction than British academic class sessions. According to Thompson (2006) and Lin (2012), while the BASE lectures are primarily monologic, the lectures of the MICASE are classified into

different levels of interactivity, namely monologic, mixed and interactive. Prior research has focused on the key role of interaction in improving learners' comprehension of academic spoken language (Chaudron & Richards, 1986; Flowerdew, 1994; Hall & Verplaetse, 2000; Morell, 2002).

The second reason is perhaps connected with the word lists used, the BNC and the BNC/COCA. Nation (2016) and Schmitt (2010) maintain that a word list reflects the nature of the corpus it has been derived from. Dang and Webb (2014) used the BNC word lists but this study used the BNC/COCA word lists. While the BNC/COCA word lists, reflecting both British and American varieties of English, incorporate very general high-frequency spoken words in the first 2000 words (Nation, 2012), the BNC word lists are strongly influenced by British nature of the BNC which is derived from mostly written corpus (made up of 90% written texts and only 10% spoken texts). Davies (2010) stated that while the BNC corpus has been useful for many different kinds of research, the COCA "is the first large, genre-balanced corpus of any language… which can be used to accurately track and study recent changes in the language" (p. 447).

The third reason for the difference between this study and Dang and Webb's (2014) findings can be related to the difference in the corpus size between the two studies and in the topics under investigation. Although the two studies analyzed the same kind of academic corpus (seminars and lectures), the corpus size and the topics examined differ greatly between the two studies. This study focused on the MICASE containing fewer lectures and seminars in comparison with the BASE. The MICASE contains 62 lectures and 7 seminars (740,651 tokens) developed from unequally-sized divisions, whereas the BASE corpus is made up of 160 lectures and 39 seminars (1,690,700 tokens) which are equally developed from four disciplinary disciplines: Arts and Humanities, Life Sciences, Physical Sciences and Social Sciences. In terms of academic written and general spoken texts, research indicates

that the coverage necessary for comprehension can vary depending on the topics studied and discourse types (Cobb & Horst, 2004; Hsu, 2014; Hyland & Tse, 2007; Nation, 2006; Webb & Rodgers, 2009). Similarly, university lectures and seminars rely heavily on topic-dependent language. For example, analyzing academic spoken texts, Dang and Webb (2014) found a large difference in the amount of vocabulary necessary to reach 95% and 98% coverage between specific disciplines, ranging from 3,000 to 5,000 word families to reach 95% coverage and 5,000 to 13,000 word families to reach 98% coverage.

It is worthwhile noting that the vocabulary sizes in the present study seem to be similar to the findings of Nation (2006) and Webb and Rodgers (2009a, 2009b) who concluded that knowledge of 3,000 word families and 6,000-7,000 word families, plus proper nouns and marginal words was necessary to reach 95% and 98% coverage of general spoken English. (see Table 9 below). This study does not claim that academic spoken English and general spoken English belong to the same type of discourse but, they both require a comparable amount of vocabulary. In fact, there are several differences between academic spoken English and general spoken English. First, in general spoken English speakers usually give speeches to large crowds and aim to reach a wider audience while the content of university lectures and seminars is delivered to specialized people involved in research in a given field. Second, the language of general spoken English does not typically contain specific scientific terminology. In contrast, academic speakers at university settings often refer to previous literature, use a particular referencing style and share findings using a systematic way (Reinhart, 2002).

Table 9

Lexical Coverage of Academic Spoken English and General Spoken English

Previous research	Type of corpus	Size (tokens)	Coverage figures (%) Vocabulary size
Dang & Webb (2014)	Academic Spoken English (BASE corpus)	1,691,997	(96.05-98) 4,000-8,000
Nation (2006)	Unscripted spoken English	200,000	(96%-98) 3,000-7,000
Webb & Rodgers (2009a)	Movies	2,841,887	(95.76%- 98.15%) 3,000-6,000
Webb & Rodgers (2009b)	TV programs	264,384	(95.45% -98.27%) 3,000-7,000
Present study	Academic Spoken English (MICASE corpus)	740,651	(95.55%-98,03) 3,000-7,000

In response to the second research question, a separate analysis of the lectures sub-corpus and the seminars sub-corpus revealed that the vocabulary size needed to reach 95% coverage is somewhat similar in these two sub-corpora. A learner needs to know 3,000 word families, plus proper nouns and marginal words, to reach 95.30% and 96.64% coverage of lectures and seminars, respectively. However, there is a discrepancy in relation to 98% coverage between lectures and seminars. The study findings suggested that 8,000 word families were necessary for the lectures sub-corpus, while 6,000 word families were sufficient to reach the same level of coverage for the seminars sub-corpus. This indicates that lectures are likely more demanding and more difficult to understand than seminars with regard to lexical coverage.

It should be noted that there are large differences between the amount of vocabulary required to reach 95% and 98% coverage for the lectures sub-corpus and the seminars sub-corpus. For the lectures sub-corpus, the difference between achieving 95% and 98% coverage was as large as 5,000 word families. When it comes to seminars, the difference between

attaining 95% and 98% was 3,000 word families, suggesting that EAP learners will need twice the vocabulary size to move from the 95% coverage point at which they need to understand the most frequent 3,000 word families, to 98% coverage. The difference between the amount of vocabulary needed to reach 95% and 98% coverage of the lectures and seminars sub-corpora lends support to the idea suggesting that lexical demands may differ based on the different types of discourse (Adolphs & Schmitt, 2004; Dang & Webb, 2014; Webb & Rodgers, 2009).

Overall, the findings suggested that knowledge of the most frequent 3,000 word families including proper nouns and marginal words provided more than 95% coverage of academic spoken English and knowledge of the most frequent 7,000 word families including proper nouns and marginal words provided 98% coverage. Learning the most 3,000 word families is an attainable goal if 95% coverage is sufficient for comprehension. However, if 98% coverage is required for comprehension, many EAP learners may find it difficult to comprehend American lectures and seminars without additional help. EAP learners should aim for receptive knowledge of the most frequent 3,000 word families rather than 7,000 word families as the minimum vocabulary size necessary to understand academic spoken text for three key reasons. First, EAP learner often read the textbook before they attend lectures. This provides more time for learners' interaction in the classroom and helps them prepare for lectures. Second, it has been indicated that learners make use of communication strategies such as the visual cues and gestures to gain better understanding of oral discourse (Adolphs & Schmitt, 2003; Harris, 2003; Mueller, 1980; Rodgers & Webb, 2016; Vidal, 2003, 2011). This may reduce the burden of challenging vocabulary in listening comprehension (Dang & Webb, 2014). Third, although 98% happens to be the ideal coverage (Nation, 2006), it is expected that learners may gain adequate comprehension of listening texts with lower than 95% (Rodgers, 2013; Van-Zeeland & Schmitt, 2012; Webb & Rodgers, 2009).

2.14 .2 The AWL Corpus Coverage

This section aims to discuss the third and fourth research questions set in this study.

3. **What is the coverage of the AWL list in the combined lectures and seminars presented in the academic spoken English?**
4. **Is the coverage of the AWL in lectures different from that in seminars?**

In response to the third research question, the coverage of the AWL in the American lectures and seminars corpus as a whole was 3.68%. This coverage figure is consistent with Dang and Webb's (2014) findings. The AWL coverage found in Dang and Webb (2014) was marginally higher (4.41%). This difference between the two studies may be explained by the corpus size (Dang and Webb's corpus contains 1,691,997 tokens; this study's corpus is just 740,651) and the number of disciplines under examination. As the corpus in Dang and Webb's (2014) study is much larger, it includes a larger number of texts from a variety of topics taken from a variety of disciplines. Dang and Webb (2014) concluded that there were some subjects demanding higher vocabulary sizes, for example, Life and Medical Sciences and Physical Sciences. Further, it is more likely that a larger corpus in terms of tokens results in a larger number of frequent items (Coxhead, 2000).

However, the coverage of the AWL in this study (3.68%) is considerably higher than Hinck's (2003) corpus of students' oral presentations. The poor coverage provided by AWL in Hinck's (2003) study (2.4%) is perhaps because the academic corpus which was analyzed was different (presentations speech) and produced by non native English speakers (Swedish speakers). Notably, the coverage figure of the AWL in this study is quite low compared with its coverage in studies of academic written corpora, for example, 10.07% (Chen & Ge, 2007), 10.0% (Coxhead, 2000), 11.6% (Cobb & Horst, 2004), 10.6% (Hyland & Tse, 2007), 10.46% (Li & Qian, 2010), 9.06% (Martínez et al., 2009), 11.17% (Vongpumivitch et al., 2009) and 11.3% (Ward, 2009). The low coverage of the AWL in academic spoken discourse is likely

because the AWL was derived from academic written texts, which emphasizes the view suggesting that the nature of word list mirrors the nature of the corpus it was developed from (Nation, 2016; Schmitt, 2010). Furthermore, the findings pertaining to the AWL coverage from this study and the studies mentioned above clearly show the general claim indicating that the AWL provided higher coverage in written texts than that in academic spoken texts. Therefore, it can be assumed that knowledge of the AWL vocabulary provided little benefit to EAP learners in relation to comprehension of the MICASE lectures and seminars corpus.

In contrast, in this study, the general words (GSL) provided high coverage in academic spoken English. As shown in Table 5, the most frequent 2,000 GSL words comprised approximately 87% of the words in American academic lectures and seminars together. The 2,000 GSL words also provided from 87% to 89% coverage of American academic lectures and seminars, respectively. These coverage points are comparable to those reported in previous studies. For example, in the research conducted by Dang and Webb (2014) and Thompson (2006), the GSL accounted for 85% to 87% coverage of academic spoken English, which is higher than its coverage in academic written discourse, (between 65% and 86%) (Coxhead, 2000; Valipouri & Nassaji, 2013). The higher coverage of the GSL in academic spoken English compared with its coverage in academic written English may be due to the some differences between written and spoken English vocabulary (Biber, 2006) (see section 2.4).

In answer to the fourth research question, the AWL provided different coverage figures in the MICASE lectures sub-corpus and seminars sub-corpus. It provided 3.80% coverage of lectures and 3.15% coverage of seminars. It should be clear that Dang and Webb (2014) examined the coverage of the AWL in the BASE corpus as a whole and across academic disciplines rather than between lectures and seminars. In the present study, the coverage of the AWL is comparable to that reported by Thompson (2006) of his economics

lectures corpus (4.90%). The small difference between AWL coverage in the two corpora may be due to the relative corpus sizes (Thompson's BASE lecture corpus contains over 1 million, the MICASE lecture corpus is just 594,117 tokens). Another possible interpretation for this difference in the AWL coverage is perhaps because of disciplinary variation of the two lecture corpora. It was found that different word families are more frequent in particular disciplines than others (Thompson, 2006). This idea was supported by Dang and Webb (2014) who found that the AWL was not equally distributed across the BASE disciplines, ranging from 3.82% in Arts and Humanities to 5.21% in Social Sciences.

The present study found a slight variation (.65%) between AWL coverage in the MICASE lectures (3.80%) and seminars (3.15%) which might be attributed to three reasons. First, unlike seminars, which are highly interactive, lectures are classified into monologic, interactive or mixed (see section 2.5). Second, the MICASE had unequally-sized sub-corpora with each sub-corpus comprising different topics and subjects: lectures (594,117 running words) and seminars (146,534 running words). This large difference between the two sub-corpora may result in different AWL coverage. According to Coxhead, Stevens, and Tinkle (2010), in larger texts, words are likely to occur more frequently in comparison to short texts which are characterized by more variation. Third, the MICASE lectures and seminars were derived from unequally-sized sub-corpora. While lectures sub-corpus was developed from the four academic divisions: Social Sciences and Education, Physical Sciences and Engineering, Humanities and Arts and Biological and Health Sciences, the seminars sub-corpus was derived from only two divisions: Social Sciences and Education, and Humanities and Arts (see section 2.9. 3.2). Hence, taking into account that different divisions and disciplines require distinct use of academic vocabulary (Chen & Ge, 2007; Martinez, Beck, & Panza, 2009), the small variation of the AWL coverage in seminars and lectures of this study can be understandable. This view was supported by the Dang and Webb (2014). In their study, the

researchers reported a considerable difference on the coverage of the AWL between Social Sciences sub-corpus (5.21) and Arts & Humanities sub-corpus (3.82). Similarly, Cobb and Horst (2004) reported a disparity regarding the AWL coverage between Medicine sub-corpus (6.72%) and History sub-corpus (14.49 %). These findings suggest that the AWL provided significantly different coverage figures based on its discipline and because lectures and seminars of this study consist of different disciplines, the coverage of the AWL in these two genres is expected to be different.

It is important to note that both the GSL and the AWL provided lower cumulative coverage in the lectures sub-corpus than seminars sub-corpus (see Table 5). This may be due to the large proportion of the technical words in some academic domains and disciplines of lectures, which in turn suggests that learners need technical words, high-frequency words (e.g., GSL) and the academic vocabulary (e.g., AWL) to understand academic lectures. Support for this view comes from Dang and Webb (2014) who found that the AWL provided the smallest coverage in their Life and Medical Sciences sub-corpus in comparison with the other three disciplinary sub-corpora, which in their opinion, is due to the large number of technical words occurring in this sub-corpus.

2.14 .3 The Usefulness of the AWL for Language Learners

The aim of this section is to discuss the fifth research question set in this study.

5. With the knowledge of the AWL, what is the vocabulary size needed to reach 95% and 98% coverage of both lectures and seminars presented in the academic spoken English?

In answer to this research question, with the aid of the AWL, learners with knowledge of proper nouns and marginal words will need a vocabulary size of 2,000 word families to reach 94.92 % coverage of American academic lectures and seminars. Importantly, this coverage figure (94.92 %) is fairly close to 95% coverage, the point at which learners can

have adequate comprehension (see Table 8). With the aid of the AWL, learners with a vocabulary size of the most frequent 3,000 BNC/COCA word families can reach 96.86% coverage of American lectures and seminars corpus. The AWL can also help learners reach 98% coverage of this corpus, if they know the most frequent 5,000 word families. However, assuming that learners do not know the AWL vocabulary, they may need 3,000 and 7,000 word families to have 95% and 98% coverage, respectively.

Because the findings revealed that there were 19, 134, and 327 word families from the AWL 570 word families occurring at the first three BNC/COCA 1,000 word-frequency level, for learners who know the most frequent BNC/COCA 3,000 word families, the support the AWL provided for such learners was from its remaining 90 word families (see Table 8). In other words, for learners with knowledge of the most frequent 3,000 BNC/COCA word families, they may need to learn only the remaining 90 AWL word families to have adequate comprehension (95% coverage) of the American academic lectures and seminars.

At the following fourth 1,000 word frequency levels, the AWL provided less support for learners. However, despite the low coverage provided by the AWL in academic American lectures and seminars in comparison to its high coverage in academic written English, it could be useful for EAP learners because it helps them save time and effort in learning academic vocabulary, and reach good comprehension of academic spoken English.

The finding of this study is similar to Dang and Webb's (2014) results because both studies suggest that the AWL can be helpful to language learners to reach adequate comprehension (95%) based on their existing vocabulary level (see Table 10 below). As shown in Table 10, the findings of this study suggested that the AWL could help learners with a vocabulary size of 3,000 word families to reach adequate comprehension of American academic lectures and seminars (95%) which is in line with Dang and Webb's (2014) results.

However, the two studies are different with respect to the support the AWL provided to learners to reach 98% coverage. Dang and Webb's (2014) findings revealed that knowledge of the AWL was not sufficient for language learners in that it did not help them to reach 98% coverage, the point at which learners have ideal coverage of academic spoken language (Webb & Rodgers, 2009). The present study, however, found that the AWL can help learners with vocabulary size of the most frequent 5,000 to reach 98% coverage of American lectures and seminars. This difference between the two studies is perhaps due to the nature of the corpora (British versus American corpus) and the word lists used (BNC/COCA versus BNC) (see section 2.14.1).

Table 10

Potential Coverage Learners of Different Lexical Sizes may Reach with Knowledge of the AWL as Suggested by this Study in Comparison to that in Dang and Webb (2014)

levels	BASE corpus Dang and Webb (2014)		MICASE corpus	
	Cumulative coverage including proper nouns and marginal words without knowledge of the AWL	Potential coverage with knowledge of the AWL	Cumulative coverage including proper nouns and marginal words without knowledge of the AWL	Potential coverage with knowledge of the AWL
1,000	87.54	90.55	86.44	90.03
2,000	92.94	94.13	91.99	94.92
3,000	94.70	95.48[a]	95.55[a]	96.86[a]
4,000	96.05[a]	96.37	96.67	97.63
5,000	96.83	96.95	97.28	98.11[b]
6,000	97.35	97.41	97.69	98.47
7,000	97.68	97.71	98.03[b]	98.78
8,000	98.00[a]	98.02[a]	98.24	98.99

[a] Reaching 95% coverage.
[b] Reaching 98% coverage.

2.15 Conclusion

This section presents a brief summary of the findings of the current study with focus on pedagogical implications of these findings.The latter part of the section discusses the limitations of the present study as well as the implications for future research on the MICASE lectures and seminars.

2.15.1 Summary of Key Findings

The aim of this study was to examine the frequency of occurrence of academic vocabulary in American academic spoken English. More specifically, this study explored the vocabulary profile of academic spoken English as represented in the MICASE lectures and seminars corpus, the coverage of the AWL in this corpus, and the extent to which the AWL may help L2 learners improve their comprehension of academic speech. For the purposes of this study, a literature review on previous investigations of lexical coverage of English texts and the AWL coverage in academic corpora was conducted. Besides, a corpus of the MICASE lectures and seminars, with 69 transcripts (740,651 tokens) was compiled and analyzed to explore the frequency of academic words using the AntWordProfiler software.

The findings of this study showed that knowledge of the most frequent 3,000 word families including proper nouns and marginal words could provide 95.55% coverage of the MICASE lectures and seminars as a whole, and knowledge of the most frequent 7,000 word families plus proper nouns and marginal words provided 98.03% coverage. However, it should be noted that American academic spoken English requires different vocabulary size to reach 95% and 98% coverage figures for lectures and seminars, which is likely due to a genre variation. With knowledge of the most frequent 3,000 word families, learners would reach 95.30% coverage of lectures and they need a vocabulary size of most frequent 8,000 word families to

reach 98.15% coverage. In contrast, perhaps because seminars involve higher levels of interactivity, they may prove to be easier for learners to understand than lectures. In order for learners to obtain 96.64% coverage of seminars, knowledge of the 3,000 most frequent word families is needed and knowledge of only the 5,000 most frequent word families provided 97.96% coverage, which is very close to the ideal comprehension figure (98%) suggested by Nation (2006).

The analysis of the AWL revealed that its coverage in the American academic lectures and seminars (3.68%) is somewhat similar to that in other academic corpora discussed in the literature review. The analysis also showed a slight difference between the coverage of the AWL between lectures (3.80%) and seminars (3.15%), which suggests its supportive role in learners' comprehension of these two academic spoken genres.

Finally, regarding the usefulness of the AWL for EAP learners, this was answered on the basis of identifying the help it provided to the learners according to their existing vocabulary size. The findings showed that although the AWL provided low coverage in American academic spoken corpus compared with that in written corpora, it plays a supporting role in helping learners comprehend academic spoken English. For example, the findings revealed that learners with a vocabulary size of 2,000-3000 word families can reach 95% coverage of American lectures and seminars if they learn the AWL 570 word families. The more interesting finding was that instead of learning 2,000 word families at the sixth and seventh 1,000 word level, with the aid of the AWL's 570 items, learners with a vocabulary size of 5,000 word families can reach 98% coverage of American academic lectures and seminars.

2.15.2 Pedagogical Implications

The intention behind the present study was to help set a vocabulary goal for EAP instructional materials which enable learners to understand American academic lectures and seminars as well as to examine the pedagogical value of the AWL in helping learners comprehend academic spoken English. Thus, the findings of this study have some pedagogical implications pertaining to vocabulary instruction.

To begin with, the findings indicated that learners of English need to recognize a considerable amount of vocabulary, in addition to proper nouns and marginal words, to have sufficient comprehension of American lectures and seminars. Learners should be aware of the 3,000 most frequently used vocabulary to gain adequate comprehension of the lectures and seminars as a whole and separately. As shown in Table 4, the 3,000 most frequent word families, plus proper nouns and marginal words, accounted for 95.55% coverage of the MICASE lectures and seminars. Therefore, if learners do not recognize vocabulary beyond 2,000-3,000 word families, it might be difficult for them to follow lectures and seminars. Research has shown that learners are unlikely to acquire the vocabulary between 3,000-6,000 word levels without additional help (Schmitt, 2010). One way to approach this problem is to draw language learners' attention to mid-frequency vocabulary which includes words between the 3,000 and 9,000 word levels (Nation, 2013; Schmitt & Schmitt, 2014). Because these words are important for learners to gain adequate comprehension, they should be encouraged to use different learning strategies and approaches to learn this large amount of vocabulary.

Academic vocabulary, in the present study, was represented by the AWL. This study analyzed the distribution of the AWL 570 word families in the MICASE lectures and seminars corpus to examine how the AWL can help learners to comprehend academic spoken English. It

was found that the AWL coverage of 3. 68% in the corpus is comparable to the AWL coverage suggested by Dang and Webb (4.41%), but quite smaller than its coverage in academic written discourse. A small variation with respect to the presence of the AWL in the text types (lectures and seminars) was also reported which indicates the specific lexical knowledge requirements of the different discourse types. The low coverage of the AWL in academic spoken corpus compared with its coverage in academic written corpus suggests that the AWL may not significantly enhance language learners' comprehension of academic spoken English. This evidence would also support the reasoning behind the assumption calling for an Academic Spoken Word List (ASWL) (Dang & Webb, 2014; Nesi, 2002; Thompson, 2006) that reflects academic spoken English. Indeed, Dang, Coxhead and Webb (2017) have recently developed an ASWL consisting of 1,741 word families to help EAP learners improve their comprehension of different types of English spoken academic discourse.

2.15.3 Limitations of the Present Study and Recommendations for Future Research

In the current study there are several limitations that should be acknowledged. One of the limitations is that the corpus size of 62 lectures (597,117 tokens) and 7 seminars (146,534) (totalling 740,651 running tokens) may be considered as small, and it is based on the MICASE Corpus which is spoken academic American data. Thus, the generalisation of the findings may be limited to the lectures and seminars in American universities. It is important to note that Dang and Webb's (2014) study did not compare the lexical profile of lectures and seminars of British academic spoken English. Therefore, to gain more reliable and generlizable findings on the lexical demands of academic spoken English, further research needs to be carried out to analyze a corpus larger than the one used in the present study representing other regional varieties of academic spoken English. For instance, it is recommended that the vocabulary size necessary to

reach 95% and 98% coverage of the BASE lectures and seminars corpus as well as different recordings of lectures and seminars be investigated in other parts of the world where English is used as medium of instruction.

The second limitation is related to the use of Coxhead's (2000) Academic Word List (AWL) to represent academic vocabulary. Although there are other academic written word lists (e.g., Gardner & Davies's (2013) Academic Vocabulary List), the AWL was selected because it has been widely used in vocabulary research and teaching materials and because of a lack of available academic spoken English list when this study was conducted. However, since a recent ASWL has been developed (Dang, Coxhead, & Webb, 2017) based on Nation's (2012) BNC/COCA lists, it would be useful if further research replicates this study to examine the value of this list in the MICASE lectures and seminars corpus as well as in other speech events of this corpus. It is worth noting that the ASWL was not yet published at the time this study was carried out; therefore, the AWL was chosen instead.

The third limitation is the use of lexical frequency levels (1,000-25,000) to provide a profile of words that learners know at each frequency level. Although previous research has indicated that learners learn vocabulary based on its frequency (Nation, 2006; Schmitt, Schmitt, & Clapham, 2001), learning does not necessarily occur based on 1,000 word units (Dang & Webb, 2014; Webb & Chang, 2012), meaning that some students may learn certain words appearing at higher frequency levels before mastering those at lower frequency levels. Therefore, one should bear in mind that in reality the cumulative coverage figures indicating the learners' vocabulary size and the coverage points suggesting the AWL support for learners reported in this study might be different. For example, in reality the lexical threshold (95%) required to adequately comprehend American lectures and seminars might be less because in lectures and

seminars learners may make use of different cues from visual aids (Miller, 2003; Rodgers & Webb, 2016; Sueyoshi & Hardison, 2005) to understand the message. Furthermore, in many lectures and seminars there is an opportunity for asking for clarifications. Overall, the findings of this study can serve a role in English language courses adopting frequency based vocabulary instruction.

The fourth limitation refers to the use of lexical coverage as a criterion of learners' vocabulary knowledge. This study only investigated lexical coverage to identify the number of words that a learner needs to know to successfully comprehend American academic spoken English. Despite the importance of lexical coverage as a predictor of textual comprehension (Laufer & Sim,1985; Stæhr, 2009), there are other factors affecting learners' comprehension of academic spoken texts, including learners' familiarity with the topic (Laufer & Nation,1995; Schmidt-Rinehart, 1994), speakers' speech rates (Rosenhouse, Haik, & Kishon-Rabin, 2006), listeners' proficiency level (Chang & Read, 2006) and the strategies learners use for listening (Hinkel, 2006; Richards, 1990).

The fifth limitation of this study is related to the use of word families as a unit of counting. This way of counting is based on the assumption that if the meaning of a headword (e.g., *access*) is known to the learners, they should be able to comprehend the relevant derived and inflected members of a word family (*accessed, accesses, accessibility, accessible, accessing,* and *inaccessible*) when presented in context, such as speaking and reading (Bauer & Nation, 1993). Although word families may be considered as an effective counting unit for native speakers of English (Nagy, Anderson, Schommer, Scott, & Stallman, 1989), acquiring or knowing word inflections and derivatives can be problematic for learners of English (Schmitt & Zimmerman, 2002).

The sixth limitation of this study is related to the AntWordProfiler program (Anthony, 2014) employed for the academic vocabulary analysis in this study. Although this program was quite helpful in analyzing the vocabulary load of the corpus of this study, it is unable to recognize multiword lexical units, for example phrasal verbs (e.g., *look forward, come in for*). Earlier research has highlighted the fact that both spoken and written discourse consists of a large proportion of multi-word units (Adolphs, 2006; Al Hassan & Wood, 2015; Erman & Warren, 2000; Nation & Webb, 2011; Schmitt, 2010).

Finally, the study findings showed that university lectures and seminars are similar to general spoken English in terms of their lexical demand. Webb & Rodgers (2009a, 200b) and Rodgers & Webb (2011) concluded that learners might be able to reach adequate comprehension of movies and television programs if they know the most frequent 3,000 word families plus proper nouns and marginal words. Rodgers (2013) conducted an intervention study to investigate EFL learners' comprehension of television programs with differing amounts of text coverage. The researcher found that learners with knowledge of only the most frequent 2,000-word level gained adequate comprehension of television programs. Dang and Webb's research and this study suggested that knowledge of between 3,000 and4,000 word families plus proper nouns and marginal words is needed to reach adequate comprehension of academic spoken language. Future experimental research can be conducted to examine the vocabulary size actually required to comprehend academic spoken English.

2.15.4 Rational for the Next Study

In the first study of this thesis, the AntWordProfiler software (Anthony, 2014) and Nation's (2012) BNC/COCA25000 word families were used as instruments to determine the lexical profile of American academic spoken English and the coverage of the AWL vocabulary

in American academic spoken English of the two discourse types: lectures and seminars. The findings of the first study also showed the methodological limitations related to the use of these two instruments. More specifically, the AntWordProfiler is unable to distinguish multiword lexical items (e.g., *come up with*) and count them as single words. Likewise, the principle of word families as a unit of counting does not include multiword items in the word definition. Thus, some vocabulary researchers recommend using lemmas instead of word families as a more reliable measure to carry out lexical coverage research (Brown, 2013; Gardner, 2007).

Lemmas refer to "a head word and its inflicted forms and reduced [*n't*] forms" (Nation, 2013, p. 7). The use of lemmas as a counting unit is based on the assumption of learning burden (Swenson & West, 1934 as cited in Nation, 2001) which simply refers to "the amount of effort required to learn [an item]" (Nation, 2013, p. 7). For example, if the base form *run* is known, it can be assumed that the other inflected forms *runs*, *ran* and *running* will have a low learning burden. The next study uses lemmas as a unit of counting to compare the distribution of the academic words represented by the AWL and multiword items, namely phrasal verbs (see Section 3.2 below for definition and characteristics of phrasal verbs) in academic spoken English.

Chapter 3 – Study 2: A corpus-based Comparison of the PHaVE List and the Academic Word List

3. Literature Review

3.1 Introduction

This chapter reviews the literature regarding phrasal verbs. It has four sections. The first section focuses on the most important considerations in relation to the definition of phrasal verbs. The following section discusses corpus-based research on phrasal verbs. The third section describes how phrasal verbs influence the word count. The fourth section recapitulates the current sections of this chapter. The research questions of this study are outlined in the last section.

3.2 Definition of the Phrasal Verbs

In spite of being "a vigorous part of English," (McArthur, 1989, p.38), the definition of the phrasal verb has been long in ongoing debate. Various theories have been made to classify phrasal verbs. For example, while there is substantial literature on both semantic and syntactic functions of phrasal verbs (e.g., Bolinger 1971; Huddleston & Pullum 2002; Palmer, 1974; Quirk et al., 1985), other specialized research focuses only on semantic (e.g., Gorlach, 2000; McIntyre, 2002; Lipka , 1972), on syntactic (e.g., Dehé, 2002; Gries & Stefanowitsch, 2004; Mitchell, 1958; Sroka, 1972), and pragmatic aspects of phrasal verbs (O'Dowd, 1998). Gardner and Davies (2007) state that there has been a basic uncertainty among language experts over the classification and description of phrasal verbs and what "to include under this fuzzy grammatical category" (p. 341). Bolinger (1971) also admittedly states "I do not believe that a linguistic entity such as the phrasal verb can be confined within clear bounds […] being or not being a phrasal verb is a matter of degree" (p. 6). However, Gardner and Davies (2007) point out that the

differences in phrasal verbs definition seem to be insignificant as far as the ESL/EFL language learners are concerned when they maintain that "if even the linguists and grammarians struggle with nuances of phrasal verb definitions, of what instructional value could such distinctions be for the average second language learner?" (p. 341).

Researchers also disagree with respect to the terms used to refer to phrasal verbs. Among these terms such as "separable verb" (Francis, 1958), "verb-particle collocations" (Sroka, 1972), "two-word verb" (Celce-Murcia & Larsen Freeman, 1999; Meyer, 1975; Siyanova & Schmitt, 2007), and "verb-particle combinations" (Fraser, 1974), phrasal verb turns out to be quite common in English literature (McArthur, 1989; Sroka, 1972) and among scholars looking at this language form (e.g., Darwin & Gray, 1999; Gardner & Davies, 2007; Garnier & Schmitt, 2015; Liao & Fukuya, 2004; Liu, 2011). Thus, for purposes of the current study, phrasal verb will be used as it is also predominant in learners' grammar textbooks and instructional materials.

However, despite the inconsistency in the description of phrasal verbs, there is a widespread acceptance shared by many linguists that phrasal verbs constitute an important part of the English language and they are often referred to as multiword verbs consisting of a lexical verb and syntactic units (an adverbial and/or prepositional particle(s)), and that their meaning varies according to the context they occur in (e.g., Bolinger, 1971; Celce-Murcia & Larsen-Freeman, 1999; Darwin & Gray, 1999; Fraser, 1976; Garnier & Schmitt, 2015; Kilby, 1984; Palmer, 1968; Quirk et al., 1985). In the current study, Gardner and Davies' (2007) definition of phrasal verb was adopted. This definition views phrasal verb as "all two-part verbs consisting of a lexical verb proper followed by an adverbial particle that is either contiguous (adjacent) to that verb or non-contiguous (i.e., separated by one or more intervening words)" (p. 341). The reasons for adopting Gardner and Davies' (2007) functional definition of phrasal verb are twofold. First,

this definition is somewhat exhaustive, especially from syntactic perspective, because it encompasses all cases of verb-particle combinations and makes no distinction on the basis of the semantic transparency of the phrasal verbs. In other words, it involves all the varying degrees of idiomaticity, including literal (e.g., *run around, run in a circle around*), idiomatic (e.g., *work out)* and any degrees of adjacency in between (*work the problem out*). As such, this definition matched the primary aim of this study that is to explore the frequency of occurrence of phrasal verbs consisting of any two-word verb + particle construction in the target corpus. The second reason for choosing Gardner and Davies' (2007) definition is that the phrasal verbs list to be investigated in this study is based on Garnier and Schmitt's (2015) research, which in turn relied on the definition of phrasal verb established by Gardner and Davies (2007).

3.3 Corpus-based Frequency Lists of Phrasal Verbs

This section deals with the four corpus-based studies on phrasal verb lists: Biber et al. (1999), Davies and Gardner (2007), Liu (2011) and Garnier and Schmitt (2015). These studies are given focused attention because they provided useful information with regard to the PHaVE List which was used as a reference list for this study. These studies are discussed in turn.

3.3.1 Biber et al's. (1999) Frequency List

Over the past few decades, there has been a growing interest in corpus-linguistic analysis of language forms, especially phrasal verbs as they are one of the most complicated aspects of English language for ESL/EFL learners (Dagut & Laufer, 1985; De Cock, 2005). Biber et al. (1999) for instance, in quantitative analysis looked at the most frequent verbs that associate with adverbial particles (AVPs) to constitute phrasal verbs in seven semantic elements (activity transitive, activity intransitive, mental transitive, occurrence intransitive, copular, aspectual intransitive and aspectual transitive). The authors also examined how relative occurrences of the

most frequent phrasal verbs vary across a number of registers: conversation, fiction, news and academic texts. Their cut-off for inclusion of a phrasal verb in their list was that it must "occur over 40 times per million words in at least one register" (p. 410) in the 40 million word Longman Spoken and Written English Corpus (LSWE). Biber et al. (1999) identified the most common 31 phrasal verbs. Their study showed that phrasal verbs are frequent in conversation, fiction and news. They are over twice as frequent in English conversation and fiction as in academic writing (roughly 2% in general conversation and 1% in academic prose) (Biber et al., 1999). This finding supports Fletcher's (2005) claim that phrasal verbs are present in written and formal English discourse where their use seems more appropriate in expressing thoughts. However, one limitation of Biber et al's (1999) findings is the limited number of phrasal verbs they came up with (31 in total). According to Gardner and Davies (2007), this limitation is probably due to space constraints as well as broader purposes and foci of their comprehensive book, the Longman Grammar.

3.3.2 Gardner and Davies' (2007) Frequency List

Gardner and Davies (2007) conducted a corpus-based study on English phrasal verbs based on the 100-million-word BNC. The key purposes of their study were to identify the actual occurrences of lexical verbs-adverbial particles combinations used in English phrasal verbs construction and to determine the potential meaning senses of each of the most prolific phrasal verbs. The corpus was analyzed and divided into two-, three-, four-, up to seven-word chunks. Then, these chunks were processed using software to identify and report all the instances where verb-plus-particle combinations occur. Next, the outcomes were lemmatized to group all inflectional forms of the same verb together such as *look*, *looked*, *looking*. The findings from the study showed that a combination of a small set of 20 lexical verbs (*go, come, take, get, set, carry,*

turn, bring, look, put, pick, make, point, sit, find, give, work break, hold, and *move*) with the 8 most frequent adverbial particles (*out, up, on, back, down, in, off,* and *over*)) account for 53.7 percent of all the 518,923 phrasal verbs that were derived from the BNC. The researchers also identified 100 high-frequency phrasal verb lemmas (e.g., *carry out, go back, pick up*), which account for 51.4 percent of all phrasal verb occurrences in the corpus. Using WordNet (Miller, 2003), they found that these 100 most frequent phrasal verbs have approximately 559 meaning senses or an average of 5.6 meanings per phrasal verb. Considering the approach this research followed to come up with the list of the most frequent 100 phrasal verbs, this list can be useful as it identifies the phrasal verbs that learners are more likely to know. However, Liu (2011) provided a number of limitations of Gardner and Davies' (2007) study. First, 100 high-frequency phrasal verbs came only from the most prolific 20 lexical verbs (e.g., *go*, *get*, etc.), as a result, other high frequent phrasal verbs may not be on the list. Second, because the study used the BNC as the only source of data, a question arises whether these phrasal verbs are also frequent in other major English varieties. Third, the study did not explore the frequently used phrasal verbs across other major registers, for example, spoken versus written English.

3.3.3 Liu's (2011) Frequency List

Liu (2011) conducted a multi-corpus analysis to compare the most frequently used phrasal verbs between American and British English discourse and to examine their usage across registers in American English. He based his analysis on data from the COCA as the primary corpus and the BNC for comparison purposes. The 40-million-word LSWE corpus was also employed for cross-corpora comparison. The researcher identified the most frequent phrasal verbs of British and American English varieties and the outcomes were analyzed statistically (i.e., chi-square and dispersion tests). The findings indicated an initial list of 152 phrasal verbs in

American and British English, which includes the phrasal verbs from Biber et al.'s (1999) and Gardner and Davies' (2007) lists. It is important to note that no significant difference was revealed between the COCA and the BNC in spite of the difference in time period covered by the two main corpora, suggesting "that PV [phrasal verb] use has remained fairly stable" (Liu, 2011, p . 671). However, the researcher indicates that there is a difference in the use of *around* and *round* in the phrasal verbs between American and British English varieties. While Americans prefer *around,* British favor *round.* Because some lexical verbs are synonymous, Liu (2011) combined them together (*look around* with *look round,* and *turn around* with *turn round,*) resulting in reduction of his final list from 152 to a total of 150 most common phrasal verbs.

Overall, the corpus-based studies mentioned above revealed an interesting and important list of the 150 most frequent phrasal verbs which learners are most likely to encounter. Although this list can be useful for teachers and learners to increase their awareness on the most frequent English phrasal verbs, Garnier and Schmitt (2015) point out that those 150 phrasal verbs have a wide variety of meanings, and thus they are overwhelming for language learners.

3.3.4. Garnier and Schmitt's (2015) Phrasal Verb Pedagogical List (PHaVE List)

Garnier and Schmitt (2015) carried out corpus-informed research based on the COCA corpus with the specific aim of reducing the overall number of meaning senses of the 150 most frequently used phrasal verbs in English, involving those suggested by Biber et al. (1999), Gardner and Davies (2007) and Liu (2011). Looking at the semantic aspect of these phrasal verbs, the researchers note that they are highly polysemous. For example, the meanings obtained from Gardner and Davies' (2007) and Liu's (2011) lists cover between 559 and 840 meanings respectively, suggesting that learners may find it difficult to distinguish between some of their overlapping senses. Using the COCA corpus, Garnier and Schmitt (2015) derived the meaning

frequency percentages of the most frequent 150 phrasal verbs. The defining criterion for inclusion of meaning senses in their list was the lower and upper frequency threshold (10% and 75% coverage) depending on the frequency of occurrence of the primary meaning of each phrasal verb. In other words, if the 75% threshold was not met by the primary sense of a given phrasal verb, additional meanings were added until the 75% or at least 10% coverage of meaning senses was reached. In their study, Garnier and Schmitt (2015) developed the PHaVE List consisting of 150 most frequent phrasal verbs along with their key meanings. A percentage number was provided for each phrasal verb meaning sense, indicating the approximate frequency of this meaning. A brief definition of these phrasal verbs and example sentences were also provided. Below is an example (Garnier & Schmitt, 2015, p. 658):

28. TAKE OFF

1. Remove STH (esp. piece of clothing or jewellery from one's body) (41%)

    I took off my shirt and went to bed.

2. Leave or depart, especially suddenly (28.5%)

    They jumped into the car and took off.

3. Leave the ground and rise into the air (14%)

    The plane took off at 7am.

Because Garnier and Schmitt's (2015) PHaVE List is the most recent and exhaustive, and was compiled and identified based on three informative studies (Biber et al., 1999; Davies & Gardner, 2007; Liu, 2011) adopting various procedures, corpora and genre types, it was considered as the basis for identifying phrasal verbs in this study. For the purposes of the current study, the coverage of the PHaVE List will be compared with that of the AWL in the target corpus.

The aim of this section was to provide a brief overview on the research investigating the phrasal verbs lists in the literature. The next section presents the case for the existence of phrasal verbs and their possible effect on learners' discourse comprehension.

3.4 The Effects of Phrasal Verbs on Word Counts

One of the key areas where corpus Linguistics has had great effect is identifying the presence of lexical collocation (Sinclair, 1991). Collocations are two or more words that co-occur more frequently than their individual frequencies within a particular discourse (Durrant, 2009; Hoey, 1991; Jones & Sinclair, 1974). Lipka (1972) calls phrasal verbs "collocations" in that "a simplex verb collocates with a particle" (p. 74). Some forms of collocations are relatively fixed and frequently referred to as formulaic sequences (Schmitt, 2004; Wood, 2007) or multi-word items (Moon, 1997). A multi-word item involves a sequence of two or more words which "semantically and/or syntactically forms a meaningful and inseparable unit" (Moon, 1997, p. 43). According to Moon (1997), this definition includes compounds (e.g., *collective bargaining*), phrasal verbs (e.g., *break off, write down*), idioms (e.g., *spill the beans*), fixed phrases (e.g., *of course*) and prefabs (e.g., *that reminds me*).

More importantly, these multi-word items consist of more than one word form and they occur frequently in spoken and written English. Erman and Warren (2000) estimate that multi-word items make up 58.6% of spoken English and 52.3% of written English. Of particular interest of this research is phrasal verbs which are highly frequent and polysemous. As pointed out earlier, for instance, examining the most frequent phrasal verbs in the BNC, Gardner and Davies (2007) found that the 100 phrasal verbs they investigated accounted for 559 potential meaning senses (an average of 5.6 per phrasal verb) ranging from transparent to idiomatic. Such

figures suggest that the estimates that depend on frequency alone for counting single word forms may be somewhat misleading for text comprehensibility.

As indicated in the first study (see section 2.6.2), current studies claim that knowledge of 2,000-3,000 word families plus proper nouns accounted for 95% coverage while between 6,000 and 7,000 word families plus proper nouns account for 98% coverage of general spoken English (Dang & Webb, 2014; Hirsh & Nation, 1992; Hu & Nation, 2000; Laufer, 1991; Nation, 2006; Webb & Rodgers, 2009). This research has used the word families included in particular lists (e.g., Coxhead, 2000; Nation, 2006; West, 1953) to estimate word knowledge. Nation (2001) suggests that once learners exceed a particular vocabulary threshold, their comprehension of texts will increase. Although this claim is true to a certain degree, the question lies in what exactly constitutes a word (Gardner, 2007). Previous corpus-based studies have shown that some words associate with other words and form distinct meaning senses (Nattinger & De Carrico, 1992; Pawley & Syder, 1983). For example, one can clearly notice the syntactic and semantic distance between the following monosyllabic verbs and their combination with prepositional or adverbial particles (phrasal verbs): *break*, *break down*, *break up;* → *get*, *get on*, *get off*; → *take*, *take off*, *take on*; → *come*, *come over*, *come across.* A learner of English is most likely to encounter these phrasal verbs in spoken and written texts. However, it is unlikely that L2 learners would need ‘‘little or no extra effort’’ (Bauer & Nation, 1993, p. 253) to recognize and understand the meaning of *break down*, and *break up*, for example.

Although Nation (2006) acknowledges the limitation of word counts based on word lists which rely on word families as a counting unit, he did not consider it a problem for language learners’ comprehension. Nation (2006) assumes that learners could know the meaning of multi-word expressions if they are non-idiomatic and transparent in meaning, and because there is a

small number of truly idiomatic phrases in English, they are ‘‘not a major issue’’ (p. 66). This logic somewhat applies to phrasal verbs in the sense that many of them are non-idiomatic in nature, suggesting that learners may deduce the meaning of phrasal verbs if the verb element is known. For example, if the learner knows *to go* or *to get*, he/she may not have a real problem in understanding *to go out* or *to get off*. However, it is debatable how easy it is for L2 learner of English language to accurately deduce the meaning of transparent phrasal verbs. Furthermore, there are phrasal verbs that exist somewhere between idiomaticity and non-idiomaticity (e.g., *eat out*) and it is unlikely that every learner is able to guess their meaning, especially in spoken context. Definitely, phrasal verbs vary in the degree of semantic opaqueness and their idiomaticity will fall on Lewis's (1993) idea suggesting that "the transparency of idiomatic expressions is a matter of degree" constructing a ‘‘spectrum of idiomaticity’’ (p. 98). In addition, Nation's (2006) argument that the idiomatic expressions (e.g., phrasal verbs) are small in number is questionable.

The studies cited above have shown that the number of frequently occurring multi-word expressions in English texts, including phrasal verbs, is higher than previously believed. Clearly, definition of vocabulary is not limited to single words, and L2 learners' comprehension of the different text types, may be not only the result of individual words, but also of the comprehension of multi-word items, including phrasal verbs. In order to investigate the influence of phrasal verbs on discourse comprehension, an assessment of their lexical coverage in different text types is needed. The primary concern of this study is academic spoken discourse. However, no attempts have been made to determine the phrasal verbs frequency in academic spoken English texts.

3.5 Summary

The above literature review has revealed that no universal definition of phrasal verbs has been reached, especially in terms of which forms constitute a phrasal verb (Biber et al., 1999; Sawyer, 2000). One of the reasons for this lack of agreement is that while some scholars define the phrasal verb as the combination of a main verb and an adverbial particle or a preposition, others only consider the phrasal verb as a main verb followed exclusively by an adverbial particle. However, phrasal verbs have traditionally been defined as a structure that consists of a main verb (such as *get*, *give*, *go*, *make*) and an adverbial particle (such as *up*, *down*, *in*, *out*). As regards to the semantic aspects of phrasal verbs, these may range from directional or spatial, transparent or literal (e.g., *take away*, *stand up*) to aspectual (e.g., *eat up*, *burn down*) to non-compositional, idiomatic or opaque, (e.g., *make up*, *look out*) (Bolinger, 1971; Celce-Murcia & Larsen Freeman, 1999; Makkai, 1972).

The survey of the above research has also shown that a considerable number of studies have examined the vocabulary size required to reach 95% and 98% coverage of written and general spoken texts (e.g., Dang & Webb, 2014; Nation, 2006; Stæhr, 2009; Van-Zeeland & Schmitt, 2013; Webb & Rodgers, 2009). There is clearly a general neglect of phrasal verbs when lexical coverage based on known words is calculated. Perhaps the lexical coverage of a discourse is overestimated when phrasal verbs are not taken into account. These studies raise the question whether variables such as the phrasal verbs can confirm or revise these vocabulary coverage and size figures. There are good reasons to expect that the frequent occurrence of phrasal verbs may be a strong determiner of lexical coverage because, so far, no other studies have been carried out to look at what impact phrasal verbs might have on such coverage estimates. Thus, comparing the coverage of the PHaVE List and the AWL in academic spoken English is warranted.

3.6 The Present Study

It is worth remembering that the coverage of the whole AWL 570 word families was examined in the first study of this thesis. This study intends to examine the coverage of the phrasal verbs represented by Garnier and Schmitt's (2015) PHaVE List in academic spoken corpus (the BASE and MICASE corpora). The purpose is to assess the value of this list for instructional purposes. To the best of the researcher's knowledge, it is the first study investigating this issue. The current study also compares the coverage of the PHaVE List with that of Coxhead's (2000) AWL in the same corpus. It is important to remember that the PHaVE List was made based on verb lemmas (base form and inflectional affixes) while the AWL was generated based on word families (base form plus the inflected and derived forms). Schmitt (2010) maintains that in order to allow for valid comparison and interpretation of results, the comparison of word lists should be based on the same unit of counting.

In order to provide an equitable comparison between the AWL and the PHaVE List in this study, the two lists should contain the same numbers of items and involve the same counting units. Due to this methodological difference in the unit of measurement, it is impossible to convert the phrasal verbs of the PHaVE List to word families. However, it is possible to convert the word families (i.e., verbs) in the AWL into lemma verbs. Therefore, the top 150 AWL word families were converted into lemma verbs to allow for valid comparison against the 150 phrasal verbs on the PHaVE List.

The AWL as a whole contains 570 word families that are organized into 10 sub-lists based on their frequency ranking, meaning that the words in sub-list one occur more frequently in academic texts than the words in the other nine sub-lists and the words in the second sub-list represent the second highest frequency words and so forth (Coxhead, 2000). For the purposes of

this study, the first 150 AWL word families, which were made from the first five sub-lists, were identified and arranged in 150 lemma verbs. To be specific, the first 150 verbs with different spellings of the same verb, such as “analyse” and “analyze” or “categorise” and “categorize” were chosen and converted to lemmas. According to the frequency criteria, these 150 lemmatized verbs should be the most frequent words in the AWL, and therefore, they should represent the most important words in the list. With this in mind, the lexical coverage of the150 phrasal verbs on the PHaVE List and the top 150 AWL lemmatized verbs in academic spoken English will be examined by the following research questions:

Research Question 1

a) What is the coverage of the PHaVE List in academic spoken English?

b) Is there a difference in its coverage in academic speech between lectures and seminars sub-corpora?

Research Question 2

a) What is the coverage of the 150 AWL lemmatized verbs in academic spoken English?

b) Is there a difference in the coverage of the 150 AWL lemmatized verbs in academic speech between lectures and seminars sub-corpora?

Research Question 3

a) Which list provides higher coverage in academic spoken English?

b) Which list provides higher coverage in academic speech from lectures and seminars sub-corpora?

3.7 Methodology

This study applies a corpus-based approach to identify the academic vocabulary and phrasal verbs in academic spoken English represented by the BASE and MICASE lectures and seminars collections. This approach was discussed in detail in section 2.9.1 of Study 1. In this section, the method of the present study is presented, which includes information about the study materials, corpus preparation, analysis and procedure.

3.7.1 Materials

The corpora selected for this study were the BASE and MICASE. These corpora are derived from only academic English spoken in university settings. This section provides information about these corpora and the materials used to analyze them.

3.7.1.1 Corpus of the Study

The corpus of the current study was made up of text files drawn from the MICASE and BASE collections. The MICASE corpus consists of 1.8 million tokens of transcripts of academic spoken American English representing a total of 152 speech events. These speech events include (Simpson-Vlach & Leicher, 2006) (the MICASE corpus was discussed fully in Study 1):

> Advising Sessions (2), Dissertation Defenses (4), Colloquia (Public Lectures) (14), Interviews (3), Discussion Sections (9), Lab Sections (8), Lectures (Small and Large) (62), Service Encounters (2), Meetings (6), Student Presentations (11), Office Hours (14), Study Groups (8), Seminars (7) and Tours (Campus/Museum) (2). (pp. 10-13)

The BASE collection, which was recorded at the universities of Warwick and Reading between 2000 and 2005, is composed of 1.6 million tokens taken from 160 lectures and 39 seminars. The corpus was created to represent four disciplinary groups with 40 lectures and 10 seminars in each domain (for details about the BASE corpus see

https://warwick.ac.uk/fac/soc/al/research/collections/base).The structure of the BASE is comparable to that of the MICASE in terms of corpus size and the four academic domains containing more or less equal data, as can be seen in Table 11.

Table 11

Statistics Regarding the Academic Domains of the MICASE and BASE Corpora

MICASE	Tokens	BASE	Tokens
Humanities and Arts	450,348	Arts and Humanities	439,110
Biological and Health Sciences	325,456	Life and Medical Sciences	433,205
Physical Sciences and Engineering	358,776	Physical Sciences	344,361
Social Sciences and Education	404,668	Social Sciences	458,989
Total	1,539,248		1,675,665

The MICASE and BASE corpora were chosen in this study for two reasons. First, these two corpora are representative of both North American (the MICASE) and British (the BASE) varieties of English, and thus they are appropriate to examine the linguistic features of academic spoken English. Second, the two corpora are entirely composed of academic spoken English and contain transcriptions of lectures and seminars in separate files, which make them correspond with the actual academic language used in university contexts. Hence, the nature of the MICASE and BASE corpora and the way in which they are organised make them the best option currently available to answer the questions of this study.

3.7.1.2 The PHaVE List and the AWL

The 150 phrasal verbs on the PHaVE List and the 150 AWL lemmatized verbs were the lists compared in the current study. It is important to remember that the PHaVE List was made

based on verb lemmas while the AWL was generated based on word families.To make a valid comparison between the the150 phrasal verbs on the PHaVE List and the AWL, the first 150 verbs of the AWL were lemmatised and organised in a lemma verbs sub-list. A list of the 150 AWL lemmatized verbs is given in Appendix B.

3.7.2 Corpus Preparation

As outlined in the previous section, the BASE corpus (approximately 1, 675,665 tokens) is quite similar in size to the MICASE (approximately 1,539,248 tokens). However, while the MICASE includes fewer examples of 15 different types of academic speech events (see section 3.6.1.1), the BASE corpus includes many more examples of only two academic speech types: lectures and seminars. The present study looks at the lectures and seminars components of the BASE and MICASE collections.

With respect to data cleaning, it is worth noting that although the BASE texts are freely available online in two formats, XML and plain texts, they had to be treated properly in order to enable effective software analysis of these texts. As such, the texts files were first extracted from their original plain texts and saved as separate Word (.doc) files. Then, these files had to go through format conversion and cleaning processes similar to that of the MICASE in order to be compatible with the corpus software. For more information about data cleaning see section 2.10.1 of Study 1.

After cleaning the BASE texts and saving them in text format, it was deemed appropriate to merge the cleansed seminars files of the two corpora (the BASE and MICASE) in a single file and the same was also done with the lectures. As a result, the corpus selected was composed of three text files labelled: BASE&MICASE_ lectures sub-corpus, BASE&MICASE_ seminars sub-corpus and the third file, BASE&MICASE lectures & seminars corpus, containing the first

and second text files together (lectures and seminars). As shown in Table 12, the final version of the corpus compiled for the present study consists of three texts from the BASE and MICASE corpora, comprising 2,431,351 tokens.

Table 12

Numbers of Words in the Two Text Types of the BASE and MICASE Corpora

	Text types	Numbers of words
1	Lectures	1,847,889
2	Seminars	583,462
3	Lectures and seminars	2,431,351

3.7.3 Corpus Analysis

This section is comprised of information about the analytical tools employed in data analysis. There are two corpus software tools used in this study. The first one is AntWordProfiler (Anthony, 2014) which is used for investigating the coverage of the 150 AWL lemmatized verbs in the target corpus. This software was presented in Study 1, section 2.11.1. The second software is AntConc (Anthony, 2014) which will be outlined below.

3.7.3.1 AntConc Software

AntConc is one of the most widely used corpus linguistic analyzing tools developed by Dr. Laurence Anthony, for English corpus processing (see Anthony, 2006 for a description of the development of the AntConc software). This software is freely available online at http://www.laurenceanthony.net/software/antconc/. AntConc provides users with different ways of accessing digital corpora through its user-friendly interface which includes a number of functions and features, such as chunking, collocation, concordance, and clusters. The Concordance tool is of central interest to this study as it displays the repeated instances of

phrasal verbs in the order of descending frequency, and helps highlight their patterns in the corpus. AntConc as a corpus analysis tool has been used by previous research involving academic use of corpus texts (Charles, 2014; Csomay & Petrović, 2012). AntConc 3.4.4w (Anthony, 2014) has been used in the present study in order to calculate the frequency counts of 150 phrasal verbs the PHaVE List from the compiled corpus.

3.8 Procedures of Corpus Analysis

The corpus in this study was analyzed in two stages specified below. The first stage describes how to explore the coverage of the 150 AWL lemmatized verbs in the corpus. The second stage outlines the steps of examining the coverage of the 150 phrasal verbs in the same corpus.

Concerning the coverage of the 150 AWL lemmatized verbs in the corpus, this was obtained according to the following procedure: The BASE&MICASE_ lectures & seminars corpus file containing lectures and seminars corpus of the BASE and MICASE was loaded as one file onto the AntWordProfiler program with the 150 AWL lemmatized verbs serving as the base word list in the program. The AntWordProfile provides the percentage coverage of the 150 AWL lemmatized verbs in the loaded corpus. To obtain the coverage of the 150 AWL lemmatized verbs in the other two corpus files (BASE&MICASE_ lectures sub-corpus, BASE&MICASE_ seminars sub-corpus), the same procedure was followed.

As far as the phrasal verbs are concerned, the major methodological issue in this study is related to their definition. As indicated in section 3.2, many researchers regard only idiomatic verb-particle combinations as phrasal verbs (e.g., Quirk et al., 1985). However, in many cases, a clear-cut differentiation between literal and idiomatic or figurative phrasal verbs is an unfeasible task partly due to their polysemy (see section 3.3). In addition, the PHaVE List generated by

Garnier and Schmitt (2015) was used as a reference for this study. Garnier and Schmitt's (2015) research relied on Liu's (2011) and Gardner and Davies' (2007) definition of the phrasal verb. These studies analyzed phrasal verbs irrespective of their semantic transparency. For these two reasons, phrasal verbs with both literal and idiomatic or figurative aspects were considered in the current study. Hence, the constructions: "*He looked up the unknown words in the dictionary*" and "*He looked up at the clouds*" were both considered.

Because one aim of the present study is to investigate the coverage of the PHaVE List in the target corpus, this list should be subject to proper treatment prior to proceeding with the analysis using AntConc software. First, all the 150 phrasal verbs in the PHaVE List were copied to a Word (.doc) file. Second, analyzing each individual phrasal verb on the PHaVE List in the study corpus is tedious and very time-consuming. Therefore, the 150 phrasal verbs on the PHaVE List were lemmatized and analyzed together. Each verb with a preposition and/or a particle in the PHaVE List was manually lemmatized to come up with its different forms (for example, the phrasal verb lemmas *give up, gives up*, *giving up, gave up, given up* for the phrasal verb lemma *give up*). A lemmatized list of all phrasal verbs on the PHaVE List is provided in Appendix A. Finally, the lemmatised PHaVE List was saved in a text file for the software analysis.

After lemmatizing the 150 verbs of the PHaVE List and saving them in plain text (txt) format, their frequency of occurrence in the compiled corpus was examined using AntConc Concordance Tool. This requires several systematic steps. First, the BASE&MICASE_ lectures & seminars corpus file containing the BASE and MICASE lectures and seminars corpus was loaded as one file onto the AntConc software. Second, the lemmatised PHaVE List was loaded on the software. AntConc offers the 'Advanced' tool next to the search box which is used to

search for phrases and sets of words. This tool was used to upload the lemmatised PHaVE List. Once this tool is clicked, a dialogue box would appear. In the resulting dialogue box (see Figure 1), the features "Use search term (s) from list below", "Load File" and "Apply" need to be consecutively selected.

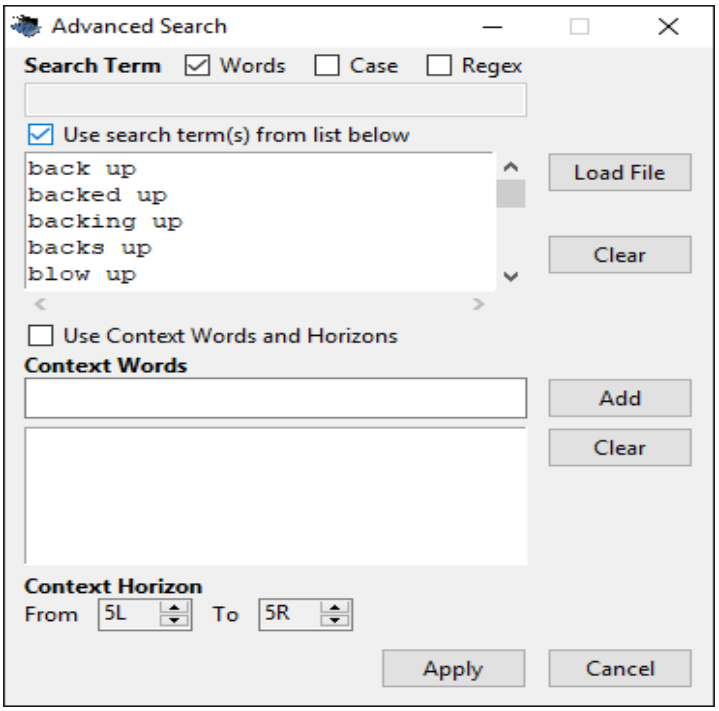

*Figure 1.*Snapshot of AntConc (version 3.4.4w for Windows) with Advance Search tool selected (after loading the lemmatized PHaVE List)

Finally, to run the search, the 'Start' feature on the AntConc screen was selected. The concordance lines display the frequencies of the 150 phrasal verbs in the PHaVE under the feature "Concordance Hits' box. Part of the concordance of the 150 phrasal verbs of the PHaVE List in the BASE&MICASE_ lectures & seminars corpus file is shown in Figure 2. As can be seen in the figure, 'Concordance Hits' box shows that there are altogether 12,826 occurrences of phrasal verbs in the BASE&MICASE_ lectures & seminars corpus file which was loaded. In order to conduct a separate analysis of the phrasal verbs frequencies of the PHaVE List in the

other two corpus files (BASE&MICASE_ lectures sub-corpus, BASE&MICASE seminars sub-corpus), the same procedure described above was followed.

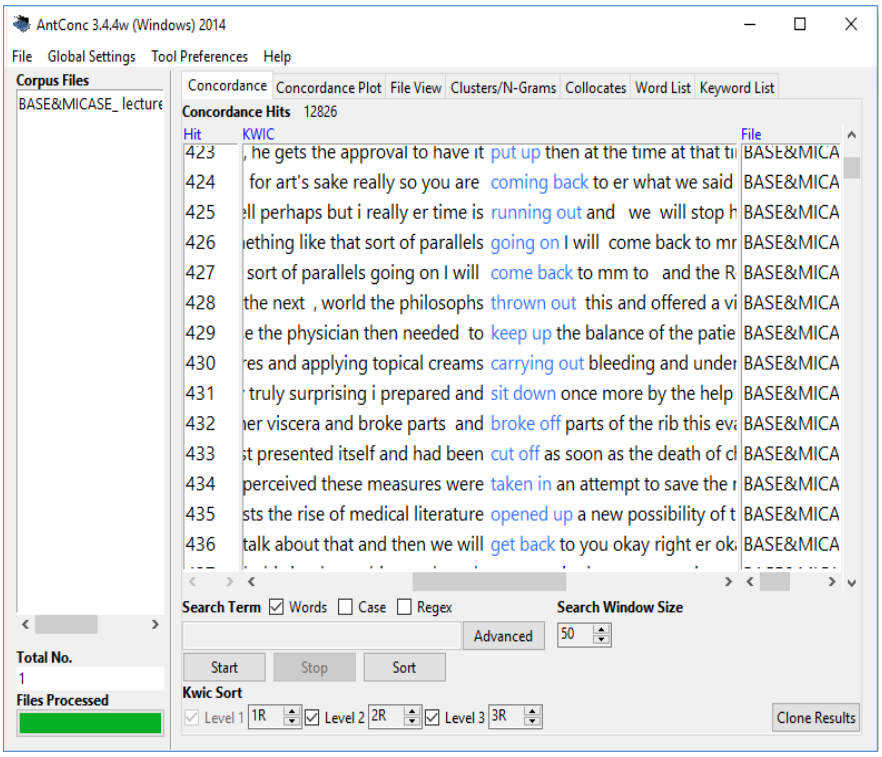

*Figure 2.*AntConc concordance lines of the loaded lemmatized PHaVE List in the BASE and the MICASE lectures and seminars corpus

Another issue to consider when searching for the phrasal verbs in this reference corpus is to find the frequencies of separable phrasal verbs which allow for intervening words to be inserted inside the phrasal verb compound. Given that past research into the phrasal verbs (Gardner & Davies, 2007; Garnier & Schmitt, 2015; Liu, 2011) looked at phrasal verbs separated by two syntactic units maximum (e.g., *turn* the device *off*), in this study the search was also limited to phrasal verbs separated by two syntactic units maximum. Because the AntConc software is unable to analyze all separable phrasal verbs on the PHaVE List, querying for them in the reference corpus was a challenging task. To examine the separable phrasal verbs in the corpus, a concordance in AntConc is first created. This requires several steps. First, the

"Concordance" tab in the top AntConc screen was first selected and then the BASE&MICASE_ lectures & seminars corpus file was loaded by means of the AntConc 'File' menu. Then, an adverbial particle or preposition was entered in the "Search Term" box which is underneath the main window and then the 'Start' button was selected to generate concordances. Next, to easily identify the verb associates with the adverbial particle or preposition searched for, the concordance line created had to be sorted. This was done by selecting two levels to left (position to the left of the search particle) under 'Kwic Sort' and pressing the 'Sort' button. Figure 3 shows part of the particle *up* concordance with the original context words sorted alphabetically to the left by two positions (2L, 0L, 0L). As demonstrated in the figure, the left-sorted concordance highlights the phrasal verbs such as *pick* this *up*, *picking* things *up*, *pick* me *up, pick* one *up* and so forth.

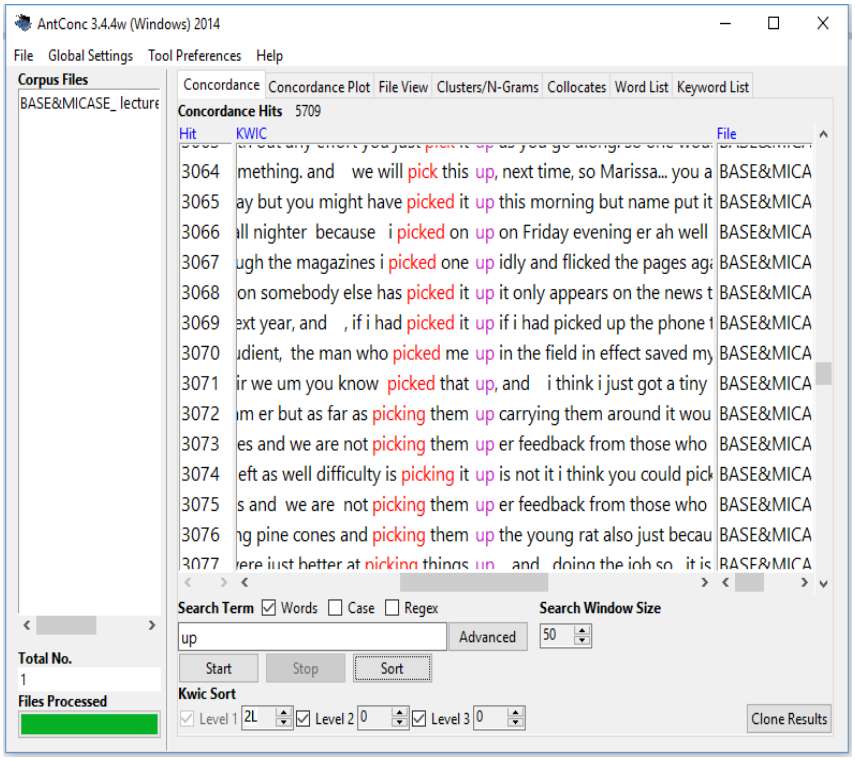

Figure 3. Part of a left-sorted AntConc concordance of the particle up in the BASE and MICASE lectures and seminars corpus.

It should be kept in mind that although the AntConc software was quite helpful in creating and organizing multiple concordance lines, it was necessary to carry out a thorough manual investigation of these concordances to identify the phrasal verbs. As shown in Figure 3, the concordance lines sorted all the inflectional forms of the verb *pick* with the particle *up* alphabetically. However, the concordances had to be examined manually further down the list to decide on the other lexical verbs associated with the particle *up* which should be considered as phrasal verbs and then counted together. This involves identifying the words that precede and follow the phrasal verbs under examination. For instance, in certain cases, a preposition combines with other words and they do not work as phrasal verb, which were then discarded. The following statement extracted from the study corpus serves as an example of this situation: "He can *get* the equipment *on* *Monday*". The preposition *on,* in this statement, does not actually fall as a unit under the phrasal verb *get on*. Rather, *get* is the verb, whereas *on* is the preposition related to the noun *Monday* (prepositional phrase).

3.9 Findings and Discussion

This section aims at presenting the findings of the corpus analysis which are obtained with the help of AntWordProfiler and the AntConc programs. These findings are organized according to the research questions. The discussion section is organized in terms of a) the coverage of the PHaVE List in the BASE and MICASE lectures and seminars corpus, b) the coverage provided by the 150 AWL lemmatized verbs in this corpus and c) comparing the coverage of the 150 AWL lemmatized verbs and the PHaVE List in this same corpus. In this section, data analysis and discussion are presented together to help answer the research questions outlined in section 3.5.

3.9.1 Findings

Research Question # 1

a) What is the coverage of the PHaVE List in academic spoken English?

b) Is there a difference in its coverage in academic speech between lectures and seminars sub-corpora?

To find the coverage of the 150 phrasal verbs on the PHaVE List in the BASE and MICASE lectures and seminars corpus, a combination of the corpus linguistic tool, AntConc and a manual linguistic analysis were employed. Table 13 shows the coverage of the 150 phrasal verbs on the PHaVE List in the whole corpus (the BASE and MICASE lectures and seminars) as well as in the two sub-corpora (lectures and seminars). As can be seen in Table 13, the BASE and MICASE lectures and seminars corpus contained a total of 2,431,351 running words (tokens), unequally divided between lectures sub-corpus (1,847,889 tokens) and seminars sub-corpus (583,462 tokens). The data in the third column represent the frequency of all the PHaVE List 150 phrasal verbs found in the corpus based on the AntConc software and manual analyses. The fifth column of Table 13 indicates the coverage of the PHaVE List in the academic spoken English corpus. The coverage provided by the PHaVE List in the corpus was calculated in the following way. First, the overall number of phrasal verbs tokens in the corpus was determined. Each of these phrasal verbs consists of two parts; verb plus particle combinations (e.g., *look up*, *sit down* and *get on)*. Therefore, to calculate the total number of phrasal verbs tokens in the corpus, it was necessary to multiply the overall raw frequency of the phrasal verbs by 2. Second, to attain the lexical coverage of phrasal verbs in this corpus the resulting figures were divided by the number of running words in the corpus and the figures obtained were multiplied by 100. For example, there were 13,198 phrasal verbs found in the combined lectures and seminars corpus.

These 13,198 phrasal verbs were multiplied by 2 (13,198 × 2 = 26,396), the figure obtained was divided by the number of this corpus tokens (26,396 ÷ 2,431,351 = 0.01085) and multiplied by 100 (0.01085× 100 =1.09%). According to this calculation result, the PHaVE List coverage found in the corpus accounted for 1.09% of the combined lectures and seminars corpus. The coverage data were presented as percentages to facilitate comparison between coverage figures of the PHaVE List in the overall corpus and the sub-corpora examined.

The coverage of the PHaVE List in the corpus varies slightly according to text type. The PHaVE List provided 1.10% coverage of lectures sub-corpus compared with 1.03% of the tokens in seminars sub-corpus. Notably, the coverage of the PHaVE List in seminars (1.03%) was almost comparable to that in the overall corpus (1.09%) despite the large difference in the tokens number between the two corpora. This is perhaps because the seminars sub-corpus used in this study contains a wide variety of seminars from different disciplines and varieties of English, and thus it is more representative of university seminars in general.

Table 13

Coverage of the 150 Phrasal Verbs on the PHaVE List in the BASE and MICASE Corpora

Text/ corpus types	**Corpus size (tokens)**	**The PHaVE List frequency**	**The PHaVE List tokens**	**The PHaVE List coverage %**
Lectures	1,847,889	10,180	20,360	1.10
Seminars	583,462	3,018	6,036	1.03
Lectures and Seminars	2,431,351	13,198	26,396	1.09

Research Question # 2

a) What is the coverage of the 150 AWL lemmatized verbs in academic spoken English?

b) Is there a difference in the coverage of the 150 AWL lemmatized verbs in academic speech between lectures and seminars sub-corpora?

To address the second question, the coverage of the 150 AWL lemmatized verbs in the different sections of the corpus was explored using the AntWordProfiler software. Table 14 shows the number of tokens of the overall corpus and per text type as well as the coverage of the 150 AWL lemmatized verbs in this corpus. Interesting similarities and differences of the coverage provided by the AWL 150 lemmatised verbs in the corpus were observed. It was found that the AWL150 lemmatised verbs provided an overall coverage of 1.00% of the running words in academic spoken English. With regard to the lectures and seminars sub-corpora, the findings suggested that the 150 AWL lemmatized verbs coverage varies considerably between these two text types. They accounted for higher coverage in lectures (1.06%) compared to seminars (0.79). However, the coverage of the 150 AWL lemmatized verbs in lectures (1.06%) is similar to its coverage in the overall corpus (1.00%).

Table 14

Coverage of the 150 AWL Lemmatized Verbs in the BASE and MICASE Corpora

Text/ corpus types	Size (tokens)	AWL 150 lemmas coverage %
Lectures	1,847,889	1.06
Seminars	583,462	0.79
Lectures and Seminars	2,431,351	1.00

Research Question # 3

a) Which list provides higher coverage in academic spoken English?

b) Which list provides higher coverage in academic speech from lectures and seminars sub-corpora?

The coverage of the 150 AWL lemmatized verbs and the 150 phrasal verbs on the PHaVE List in the overall corpus and in the two sub-corpora (lectures and seminars) is shown in Table 15. Coverage is presented as a percentage of all tokens in the corpus. The 150 AWL lemmatized verbs and the 150 phrasal verbs on the PHaVE List are similar with regard to their coverage in the overall corpus. The coverage of the 150 AWL lemmatized verbs covered 1.00% of running words in academic spoken English corpus and the 150 phrasal verbs on the PHaVE List provided coverage of 1.09% of the words in the same corpus. By contrast, the PHaVE List 150 phrasal verbs provided higher coverage in lectures and seminars sub-corpora than the 150 AWL lemmatized verbs. The PHaVE List 150 phrasal verbs covered 1.10% of lectures sub-corpus compared to 1.06% by the 150 AWL lemmatized verbs. Similarly, the PHaVE List 150 phrasal verbs covered 1.03% of the words in the seminars sub-corpus while the 150 AWL lemmatized verbs covered 0.79%.

Table 15

Coverage of the 150 AWL Lemmatized Verbs and 150 Phrasal Verbs on the PHaVE List in the BASE and MICASE Corpora

Text/ corpus types	Size (tokens)	AWL 150 coverage %	PHaVE List coverage %
Lectures	1,847,889	1.06	1.10
Seminars	583,462	0.79	1.03
Lectures and Seminars	2,431,351	1.00	1.09

3.10 Discussion

The current study aimed to determine the coverage of the 150 AWL lemmatized verbs and the 150 phrasal verbs from the PHaVE List in the BASE and MICASE lectures and seminars corpus, and then to examine which list provides higher coverage in this corpus. This discussion section outlines and clarifies the findings regarding the study's aims. Specifically, it revolves around the research questions outlined in section 3.6.

3.10.1 Coverage of the PHaVE List in Academic Speech (Research Question # 1)

The first question in the study addressed the percentage of the 150 phrasal verbs on the PHaVE List in academic spoken language as represented by the BASE and MICASE lectures and seminars corpus. The findings of the study revealed that the overall coverage of the 150 phrasal verbs on the PHaVE List in this corpus was 1.09%. Furthermore, these phrasal verbs occur more often in the lectures texts than in the seminars texts. This is clear from the findings which indicated a slight variation in the coverage of the PHaVE List in the two genres, with the higher coverage occurring in lectures (1.10%) than in seminars (1.03%). One reason for this may be due to the fact that lectures sub-corpus involved larger number of running words than seminars sub-corpus. Indeed, the number of tokens in lectures sub-corpus accounted for more than three times the token numbers in seminars sub-corpus. Research has shown that longer texts perhaps account for higher coverage of words compared to short texts which are more likely to exhibit much more lexical variation (Coxhead, Stevens, & Tinkle, 2010; Nation & Webb, 2011; Paribakht & Webb, 2016). Another reason for the different coverage figures between lectures sub-corpus and seminars sub-corpus is perhaps related to the nature of lectures examined. The lectures sub-corpus used in this study represents both British and North American varieties of English. These have dialogic and monologic features (Hyland, 2009; Lynch, 2011; Rodgers &

Webb, 2016; Thompson, 2005), consist of conversation-like and informal language (Hansen & Jensen, 1994; Swales, 2004) and capture a wide variety of disciplinary fields. Due to these multifaceted aspects of lectures, one would expect a broader coverage of the PHaVE List.

Because no prior studies have specifically examined the coverage of the PHaVE List in English discourse, there were no previous figures for comparison. However, the coverage of the PHaVE List (1.03-1.10 %) in the current study is in line with the findings of previous two studies examined the coverage of all English phrasal verbs in written and informal English discourse. Biber et al. (1999) found that the coverage of phrasal verbs was around 2% in conversation and 1% in academic prose. Similarly, the coverage of the PHaVE List found in this study is also congruent with Vilkaitė's (2016) study of phrasal verbs in different registers (conversations, fictions, and newspaper) which showed that phrasal verbs made up 0.5-1.3% of the examined corpus.

It is surprising that the findings of this study are somewhat similar to the findings of Biber et al. (1999) and Vilkaitė (2016) despite clear differences between this study and those two studies. To be more precise, while this study examined the university-based spoken English, the other two studies examined the general conversation and written discourse. Moreover, this study aimed to investigate only the coverage of the150 most frequent phrasal verbs of the PHaVE List in academic speech, whereas Biber et al.'s (1999) and Vilkaitė's (2016) studies focus was on examining the distribution of all English phrasal verbs in conversations and academic prose. Based on these differences, one would have expected that the PHaVE List coverage to be lower given the fact that it is not an exhaustive list of all English phrasal verbs. However, it can be argued that the comparable coverage figures of only the most frequent 150 phrasal verbs on the PHaVE List in academic speech and all English phrasal verbs in the corpus examined by Biber et

al. (1999) and Vilkaitė (2016) provide clear evidence on the usefulness and instructional value of the PHaVE List.

3.10.2 Coverage of the 150 AWL Lemmatized Verbs in Academic Speech (Research Question # 2)

The findings related to the second research question showed that the 150 AWL lemmatized verbs provided 1.00% coverage of the tokens in academic spoken English (see Table 14). It is worth noting that while there are many prior studies reporting the coverage provided by all the AWL 570 word families in academic written (e.g., Coxhead, 2000; Hyland & Tse, 2007) and spoken discourse (e.g., Dang & Webb, 2014), no prior studies have ever looked at the coverage of the 150 AWL lemmatized verbs in any discourse types, meaning that the findings of this study in relation to the coverage figures of the 150 AWL lemmatized verbs cannot be compared to any earlier research. Yet, given the established argument that the 150 AWL lemmatized verbs involved in the present study represent the most frequent verbs in the AWL (Coxhead, 2000), it is reasonable to expect that they are viewed as the most useful words for language learners to know. Still, the coverage of the 150 AWL lemmatized verbs (1.00%) found in this study is notably low. This is perhaps due to two reasons. The first reason refers to the corpus which the AWL was derived from. The AWL was compiled from written academic texts (Coxhead, 2000), and therefore it is representative of academic written English. However, this study explored its coverage in academic spoken texts compiled from the BASE and MICASE corpora. The second reason may be related to the nature of the 150 AWL lemmatized verbs under investigation. The coverage of the 150 AWL items in this study was based on lemma verbs rather than word families as the way of counting which significantly reduces the number of items in a corpus (Bauer & Nation, 1993; Nation & Anthony, 2013), and as result lemma would give

lower coverage as evidenced in this study. For example, the whole 570 word families in the AWL provide a total of 3,082 tokens, whereas the 150 AWL lemmatized verbs, whose coverage was investigated in this study, provide 655 tokens.

The 150 AWL lemmatized verbs provided different coverage percentages in lectures and seminars sub-corpora. They accounted for 1.06% of running words in lectures compared to 0.79% of that in seminars. The higher coverage of the 150 AWL lemmatized verbs in lectures may be explained by the nature of academic lectures. According to Biber (2006), Flowerdew and Millar (1997) and Thompson (2015), lectures are positioned somewhere between conversation and academic writing. Therefore, one would expect to find higher coverage of the 150 AWL lemmatized verbs, which are originally derived from academic written prose, in lectures sub-corpus. The different coverage figures also suggest that the 150 AWL lemmatized verbs may benefit language learners differently. For example, because the 150 AWL lemmatized verbs appear more frequently in lectures than seminars, it is possible to assume that EAP learners attending lectures have a greater opportunity to encounter academic vocabulary in comparison to seminars. In other words, one might expect that the 150 AWL lemmatized verbs will be less useful to language learners attending seminars which exhibit higher levels of discussion and interactivity compared to lectures which are more formal and informational in nature (Flowerdew & Miller, 1997; Rodgers & Webb, 2016).

3.10.3 Comparing the Coverage of the PHaVE List and the 150 AWL Lemmatized Verbs (Research Question # 3)

In answer to the third research question, the findings showed that there was a little difference between the coverage of the 150 AWL lemmatized verbs and the150 phrasal verbs on the PHaVE List in the BASE and MICASE lectures and seminars corpus as a whole (see Table

15). The findings also revealed comparable coverage of both the 150 AWL lemmatized verbs and the PHaVE List in relation to lectures sub-corpus, suggesting that these two lists are important in academic spoken English. In other words, learning the two lists may enable learners to gain more academic text coverage, which in turn leads to a better comprehension of academic spoken English.

However, focusing on the PHaVE List is probably more cost-effective in terms of time spent on learning. This list accounts for the most important high frequency 150 phrasal verbs which learners and teachers can focus on instead of being overloaded with a large number of the AWL words which might be overwhelming and difficult for learners to manage. In addition, it has been indicated that language learners avoid using phrasal verbs especially because of their semantic complexity (Cornell, 1985; Dagut & Laufer, 1985; Garnier & Schmitt, 2015; Liao & Fukuya, 2004). To handle this problem and promote learners' comprehension, the PHaVE List contains the most useful and frequent 75% of the meanings of the 150 phrasal verbs it includes, together with definitions and example sentences. Hence, one can argue that knowledge of the PHaVE List may provide greater benefit for learners with respect to comprehension of academic speech.

3.11 Conclusion

This section concludes the current study by summarizing its findings and discussing some pedagogical implications related to EAP context. In addition, it presents some limitations of the exploration conducted and suggests future research that might be developed based on the findings obtained in this study.

3.11.1 Summary of Key Findings of Study 2

The aim of this study was to investigate the coverage of academic words and phrasal verbs in academic English spoken discourse. The 150 AWL lemmatized verbs list was organized to investigate the coverage of academic vocabulary in the BASE and MICASE lectures and seminars corpus, and the 150 phrasal verbs on the PHaVE List were used to investigate the coverage of this list in the same corpus. Furthermore, this study compared the coverage of the two lists in the overall corpus (lectures and seminars together) and in lectures and seminars separately. Two computer programs, AntWordProfiler and AntConc, were used to determine the coverage of the two lists in the corpus totalling approximately 2,431,351 running words.

The findings of this study revealed some insights into the coverage of the 150 AWL lemmatized verbs and the most frequent 150 phrasal verbs on the PHaVE List. The analysis showed that the 150 AWL lemmatized verbs provided 1.00 % of the running words in combined university lectures and seminars corpus, 1.06% in lectures sub-corpus and 0.79% in seminars sub-corpus. Concerning the 150 phrasal verbs on the PHaVE List, they accounted for (1.09%) of the running words in the combined corpus, (1.10%) in lectures sub-corpus and (1.03%) in seminars sub-corpus. It is clear that the PHaVE List provided higher coverage in the examined corpus, perhaps suggesting its pedagogical advantage over the 150 AWL lemmatized verbs in helping language learners comprehend university academic lectures and seminars.

3.11.2 Pedagogical Implications

The present study investigated the coverage of two different corpus-based lists. The 150 AWL lemmatized verbs and the PHaVE List comprising the most frequent 150 English phrasal verbs. Based on the findings of this study, some pedagogical recommendations for learners, teachers and material writers can be presented.

The findings of the current study were based on corpus analysis. A great deal of research has indicated the importance of corpus analysis in identifying distinct language features (Granger & Meunier, 2008; Lindquist, 2009) and in its contribution to language teachers and learners (Boulton, Carter-Thomas, & Rowley-Jolivet, 2012; O'Keeffe & McCarthy, 2010). Although this study is not primarily concerned with teaching and learning, language learners with some guidance and support can benefit from the freely available online corpora, such as the BNC and COCA to learn the syntactic and the semantic patterns of the most commonly used phrasal verbs included in the PHaVE List. Teachers can also have access to different patterns and meanings of these phrasal verbs with the help of the concordance lines provided by online corpus-based tools, such as AntConc (Anthony, 2014). For example, teachers can generate and arrange concordance lines in such a way to highlight patterns and meanings of the most frequent and pedagogically useful phrasal verbs on the PHaVE List. Furthermore, writers of teaching and learning materials can make use of the instructional PHaVE List and include a varied selection of exercises and assessment activities on phrasal verbs in EFL textbooks based on the this list.

Moreover, earlier research into formulaic language has shown its key role in the acquisition of second language (Conklin & Schmitt, 2012; Nation & Webb, 2011; Nattinger & DeCarrico, 1992; Pawley & Syder, 1983; Wood, 2009, 2010). Formulaic language, in this study, was represented by the PHaVE List which includes the most frequent 150 phrasal verbs along with their most frequent meaning senses (Garnier & Schmitt, 2015). Data analysis of the current study provided additional evidence for the pedagogical usefulness of the PHaVE List with respect to comprehension of university lectures and seminars. The findings revealed that the 150 phrasal verbs contained in the PHaVE List occurred more often in university lectures and seminars than the 150 AWL lemmatized verbs. Even if the coverage of the PHaVE List (1.09%)

seems rather low compared to the coverage of the whole AWL 570 word families (3.68%.) in academic spoken English (see section 2.13.2 of study 1), one can draw conclusion that effective vocabulary teaching would require teachers to teach the PHaVE List as it is more time-efficient. Research has pointed out that vocabulary knowledge requires learners to master different aspects of a given word, namely form (pronunciation, spelling), meaning (form-meaning relationship) and use (grammatical functions, collocations, register) (Nation, 2001, 2013; Schmitt, 2014). Given the large number of AWL words (570 word families) compared to the PHaVE List (150 phrasal verbs) and limited class time dedicated to vocabulary teaching and learning, managing these various facets of word knowledge in terms of the AWL words will be time consuming and demanding for both teachers and learners. A better alternative would likely be the PHaVE List because teaching the 150 phrasal verb lemmas in comparison to the 570 word families would save teachers' times and efforts and may result in a more efficient and effective teaching and learning.

3.11.3 Limitations

The findings presented in this study are limited in a number of ways. For the purposes of this study, lectures and seminars of the BASE and MICASE were merged together. While this helped in looking at the presence of the 150 AWL lemmatized verbs and the PHaVE List in both North American and British varieties of English, lectures in two corpora may differ in the degree of interactivity (Thompson, 2006) and regional variation may affect the way seminars and lectures are delivered, which in turn may result in different language use. Therefore, separate analysis of the lectures and seminars in the BASE and MICASE corpora could be conducted.

Another limitation is related to the size of lectures and seminars sub-corpora. These two sub-corpora were not equally sized. This is an unavoidable limitation of the present study

because the purpose is to compare these two corpora. However, for valid comparisons, it is better to compare corpora with an equal number of tokens (Nation & Webb, 2011) because if corpora had the same number of running words, the coverage figures could be different.

3.11.4 Further Research Suggestions

Based on the findings and limitations of this study, some ideas for further investigation can be suggested. First, this study compared the coverage of the 150 AWL lemmatized verbs and the 150 phrasal verbs on the PHaVE List in university lectures and seminars texts. Dang, Coxhead, and Webb (2017) have recently developed an academic spoken word list (ASWL) consisting of 1,741 word families derived from an academic spoken corpus totalling 13,029,661 words. The researchers indicated that language learners can achieve 92%-96% coverage of academic spoken texts if they learn the ASWL. This list was not available when this study was carried out. Therefore, further research investigating the coverage of the ASWL in the BASE and MICASE corpora and other corpora would provide more reliable and accurate assessment in relation to its coverage in academic spoken English.

Second, in this study, the most frequent 150 phrasal verbs on the PHaVE List were lemmatized and their coverage was explored in academic speech. In order for the learners to attain optimal benefit from the PHaVE List, they need to learn the lemmatized verbs (e.g., *take on, takes on, took on, taking on, taken on*). It is suggested that future studies may look at how these lemmatized phrasal verbs impact language learners' actual comprehension, especially those who are at low proficiency level.

Third, the present study is theoretical in nature and it is hoped that it will contribute to a greater awareness among teachers and learners in relation to the pedagogical value of the PHaVE List. That said, a potential area for future research might be to empirically investigate how the

phrasal verbs on the PHaVE List are used and how their semantic features may influence learners' actual comprehension.

Fourth, future research may consider investigating the PHaVE List coverage in other academic texts, for example in other speech events of the MICASE collections such as discussions, dissertation defences, and interviews. This study looked at the frequency of occurrence of the 150 phrasal verbs on the PHaVE List in the lectures and seminars of the BASE and MICASE corpora. Therefore, further research investigating the coverage of the PHaVE List in different academic speech events might complement this research.

Finally, another topic that might be important is investigating the coverage of other formulaic categories in the BASE and MICASE corpora. Earlier research studied various language patterns and came up with instructional formulaic lists representing the most salient formulaic sequences in academic language (Martinez & Schmitt's (2013) PHRASE List; Simpson-Vlach & Ellis' (2010) The Academic Formulas List (AFL)). Thus, it could be interesting and pedagogically useful to look at the similarities and differences in the coverage of the 150 phrasal verbs on the PHaVE List and other categories of formulaic language in the BASE and MICASE corpora.

3.12 Final Conclusions

Using the corpus linguistic approach, this thesis provided new insights into the lexical profile of spoken academic English and the lexical coverage of the AWL in spoken academic English. Importantly, it provided evidence for the usefulness of integrating the PHaVE list into pedagogical materials and learning syllabuses for English for Academic Purposes. In so doing, this thesis established a pedagogical value for the 150 phrasal verbs on the PHaVE List for academic spoken English and provided a rationale for considering this list when examining

learners' comprehension of academic spoken English. There is also hope that the data resulting from this thesis will contribute to raising awareness among English language teachers, learners, curriculum designers and material developers of the importance of corpus data in language learning and teaching, especially to understand the learners' needs with regard to various language features, including academic vocabulary and phrasal verbs.

Acknowledgements

1. The transcriptions (and recordings) used in this study come from the British Academic Spoken English (BASE) corpus project. The corpus was developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson. Corpus development was assisted by funding from BALEAP, EURALEX, the British Academy and the Arts and Humanities Research Council http://www.warwick.ac.uk/go/base/

2. MICASE is on-line, searchable collection of transcripts of academic speech events recorded at the University of Michigan https://quod.lib.umich.edu/m/micase/

References

Adolphs, S. (2006). *Introducing electronic text analysis. A practical guide for language and literary studies*. London: Routledge.

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425-438.

Adolphs, S., & Schmitt, N. (2004). Vocabulary coverage according to spoken context. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 39–49). Amsterdam: John Benjamins Publishing.

Al Hassan, L., & Wood, D. (2015). The effectiveness of focused instruction of formulaic sequences in augmenting L2 learners' academic writing skills: A quantitative research study. *Journal of English for Academic Purposes*, *17*, 51-62.

Altenberg, B. (1998). On the phraseology of spoken English: the evidence of recurrent word combinations. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 101-122). Oxford: Clarendon Press.

Anthony, L. (2006). Concordancing with AntConc: An introduction to tools and techniques in corpus linguistics. *JACET Newsletter*, *155*, 155-185.

Anthony, L. (2014). AntConc (Version 3.4.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/.

Anthony, L. (2014). AntWordProfiler (Version 1.4. 1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software.

Appel, R., & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: Differences between high-and low-proficiency levels. *Language Assessment Quarterly*, *13*(1), 55-71.

Aston, G. (1997). Small and large corpora in language learning. In B. Lewandowska-Tomaszczyk, & J. P. Melia (Eds.), *Practical applications in language corpora* (pp. 51–62). Lodz: Lodz University Press.

Bamford, J. (2004). Gesture and symbolic uses of the deictic here in academic lectures. In K. Aijmer, & A. B. Stenström (Eds.), *Discourse patterns in spoken and written corpora* (pp. 113-138). Amsterdam: J. Benjamins.

Basturkmen, H. (1996). *Discourse in academic seminars: Structures and strategies of interaction*. (Unpublished doctoral dissertation). Aston University, Birmingham.

Basturkmen, H. (2016). Dialogic interaction. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (PP. 165-176). London: Routledge

Bauer, L. (1993). *Manual of information to accompany the Wellington Corpus of Written New Zealand English.* Wellington, New Zealand: Victoria University of Wellington.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers.* Amsterdam: Benjamin

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics, 14*(3), 275-311.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, *36*(1), 9-48.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow, England: Longman.

Bolinger, D. (1971). *The Phrasal Verb in English*. Cambridge and Massachusetts: Harvard University Press.

Bonk, W. (2000) 'Second language lexical knowledge and listening comprehension'. *International Journal of Listening*, *14*(1), 14–31.

Boulton, A. (2012). Corpus-informed research and learning in ESP: Issues and applications. In A. Boulton, S. Thomas & E. Rowley-Jolivet (Eds.) *Corpus consultation for ESP: A review of empirical research* (pp. 261-292). Amsterdam: John Benjamins.

Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics, 36*(1), 1-22.

Brown, D. (2013).Types of words identified as unknown by L2 learners when reading. *System*, *41*(4), 1043-1055.

Browne, C. (2014). A New General Service List: The better mousetrap we've been looking for. *Vocabulary Learning and Instruction*, *3*(2), 1-10.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Camiciottoli, B. C. (2004). Interactive discourse structuring in L2 guest lectures: Some insights from a comparative corpus-based study. *Journal of English for Academic Purposes*, *3*(1), 39-54.

Campion, M. & Elley, W. (1971). *An Academic Word List*. Wellington: New Zealand Council for Educational Research.

Carey, S. (1999). The use of WebCT for a highly interactive virtual graduate seminar. *Computer Assisted Language Learning*, *12*(4), 371-380.

Carter, R. (2012). *Vocabulary: Applied linguistic perspectives*. London: Routledge.

Celce-Murcia, M., & Larsen-Freeman, D. (1999).*The grammar book: An ESL/EFL teacher's reference course*. Boston: Heinle &Heinle.

Chang, A. C., & Read, J. (2006). The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly, 40*(2), 375–420.

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes*, *35*(1), 30-40.

Chaudron, C., & Richards, J. C. (1986). The effect of discourse markers on the comprehension of lectures. *Applied Linguistics*, *7*(2), 113–127.

Cheng, L., Myles, J., & Curtis, A. (2004). Targeting language support for non-native English-speaking graduate students at a Canadian university. *TESL Canada Journal*, *21*(2), 50-71.

Cheng, W. (2012). *Exploring corpus linguistics: Language in action*. London/New York: Routledge.

Chen, Q., & Ge, C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles. *English for Specific Purposes, 26*(4)*,* 502-514.

Chujo, K., & Utiyama, M. (2006). Selecting level-specific specialized vocabulary using statistical measures. *System*, *34*(2), 255-269.

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System*, *32*(2), 251– 263.

Clark, M. K., & Ishida, S. (2005). Vocabulary knowledge differences between placed and promoted EAP students. *Journal of English for Academic Purposes*, *4*(3), 225-238.

Coady, J. (1997). L2 vocabulary acquisition through extensive reading. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 225-237). Cambridge: Cambridge University Press.

Cobb, T. (2010). Learning about language and learners from computer programs. *Reading in a Foreign Language, 22*(1)*,* 181-200.

Cobb, T., & Horst, M. (2004). Is there room for an AWL in French? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 15–38). Amsterdam, the Netherlands: John Benjamins.

Cobb, T., & Horst, M. (2015). Learner corpora and lexis. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 185–206). Cambridge, UK: Cambridge University Press.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*(1), 72-89.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45-61.

Connor, U., & Upton, T. A. (Eds.). (2004). *Discourse in the professions: Perspectives from corpus linguistics.* Amsterdam: John Benjamins.

Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics 22*, 75–95.

Cook, G. (2003). *Applied linguistics*. Hong Kong: Oxford University Press.

Cornell, A. (1985). Realistic goals in teaching and learning phrasal verbs. *IRAL-International Review of Applied Linguistics in Language Teaching*, *23*(4), 269-280.

Corson, D. (1997). The learning and use of academic English words. *Language Learning*, *47*(4), 671-718.

Courtney, R. (1983). *Longman dictionary of phrasal verbs*. Harlow, England: Longman.

Cowan, J. R. (1974). Lexical and syntactic research for the design of EFL reading materials. *TESOL Quarterly, 8*(4), 389-400

Cowie, A. P., & Mackin, R. (1993). *Oxford Dictionary of Phrasal Verbs*. Oxford: Oxford University Press.

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

Coxhead, A. (2006). *Essentials of teaching academic vocabulary*. Boston, U.S.: Houghton Mifflin Company.

Coxhead, A. (2016). Acquiring academic and discipline specific vocabulary. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (PP. 165-176). London: Routledge.

Coxhead, A., & Nation, P. (2001). The specialized vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 252–267). Cambridge: Cambridge University Press.

Coxhead, A., Stevens, L., & Tinkle, J. (2010). Why might secondary science textbooks be difficult to read? *New Zealand Studies in Applied Linguistics*, *16*(2), 35–52.

Csomay, E., & Petrović, M. (2012). "Yes, your honor!": A corpus-based study of technical vocabulary in discipline-related movies and TV shows. *System*, *40*(2), 305-315.

Dagut, M., & Laufer, B. (1985). Avoidance of phrasal verbs—A case for contrastive analysis. *Studies in Second Language Acquisition*, *7*(1), 73-79.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The Academic Spoken Word List. *Language Learning*, *67*(4), 959-997.

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*, 66-76

Darwin, C. M. & Gray, L. S. (1999). "Going after the phrasal verb: An alternative approach to classification". *TESOL Quarterly*, *33*(1): 65-83.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447-464.

De Cock, S. 2005. "Learners and phrasal verbs". *Macmillan phrasal verbs plus*. Oxford: Macmillan Publishers Limited. LS 16-LS20

Dehé, N. (2002). *Particle verbs in English: Syntax, information structure and intonation.* Amsterdam, Philadelphia: Benjamins.

DeVito, J. A. (1967). A linguistic analysis of spoken and written language. *Communication Studies*, *18*(2), 81-85.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, *28*(3), 157-169.

Durrant, P. (2014). Discipline and level specificity in university students' written vocabulary. *Applied Linguistics, 35*(3), 328 - 356.

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, *26*(2), 163-188.

Dworzynski, K., & Howell, P. (2004). Predicting stuttering from phonetic complexity in German. *Journal of Fluency Disorders*, *29*(2), 149-173.

Eldridge, J. (2008). “No, there isn’t an ‘academic vocabulary,’ but…”, *TESOL Quarterly, 42*(1), 109-113.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. J*ournal of Second Language Writing*, *4*(2), 139 – 155.

Engels, L.(1968). The fallacy of word-counts. *International Review of Applied Linguistics in Language Teaching*, *6*(3), 213–231.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, *20*(1), 29-62.

Farrell, P. (1990). *A lexical analysis of the English of electronics and a study of semi-technical vocabulary*. Dublin: Trinity College.

Feak, C. B. (2012). ESP and speaking. In B. Paltridge & S. Starfield (Eds.), *The Handbook of English for Specific Purposes* (pp. 35-53): John Wiley & Sons, Ltd.

Ferris, D. (2009). *Teaching college writing to diverse student populations*. Ann Arbor: University of Michigan Press.

Field, J. (2011). Into the mind of the academic listener. *Journal of English for Academic Purposes*, *10*(2), 102-112.

Fielder, C. (2011). “An EAP course as preparation for academic study in English”. *Journal of International Association of Teachers of English as a Foreign Language ESP Special Interest Group 38*, 15-23.

Flowerdew, J. (1994) ‘Research of relevance to second language lecture comprehension: An overview’. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 7–30). Cambridge: Cambridge University Press.

Flowerdew, L. (2004). The argument for using English specialized corpora. In U. Connor & T. A. Upton (Eds.), *Discourse in the professions: perspectives from corpus linguistics* (pp. 11-33). Amsterdam: John Benjamins.

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy. *International Journal of Corpus Linguistics*, *14*(3), 393-417.

Flowerdew, J., & Miller, L. (1997). The teaching of academic listening comprehension and the question of authenticity. *English for Specific Purposes*, *16*(1), 27-46.

Folse, K. S. (2004). *Vocabulary myths: Applying second language research to classroom teaching*. Ann Arbor, MI: University of Michigan Press.

Fortanet, I. (2004). The use of "we" in university lectures: Reference and function. *English for Specific Purposes, 23*(1), 45–66.

Francis, W. N. (1958). *The structure of American English*. New York

Fraser, B. (1974). The phrasal verb in English by Dwight Bolinger. *Language*, *50*(3), 568. doi:10.2307/412224

Furneaux, C., Locke, C., Robinson, P. & Tonkyn, A. (1991). 'Talking heads and shifting bottoms: The ethnography of academic seminars'. In P. Adams, B. Heaton & P. Howarth (Eds), *Socio-cultural issues in English for specific purposes* (pp. 75-88). Oxford: Modern English Publications.

Garcia, G. (1991). Factors influencing the English reading text performance of Spanish-speaking Hispanic children. *Reading Research Quarterly, 26*(4), 371–392.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*(2), 241-265.

Gardner, D. & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly, 41*(2), 339-360.

Gardner, D. & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305 - 327.

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, *19*(6), 645-666.

Gass, S. & Selinker, L. (2008) *Second language acquisition: An introductory course.* New York: Routledge/Taylor and Francis Group.

Gee, J. P. (1990). *Social linguistics and literacies: Ideologies in discourses.* New York, NY: Falmer Press.

Goh, C.C.M. (2013) 'ESP and Listening'. In B. Paltridge, & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 55–76). Oxford: Wiley-Blackwell.

Gonzelez, M. C. (2013). *The intricate relationship between measures of vocabulary size and lexical diversity as evidenced in non-native and native speaker academic compositions* (Unpublished doctoral dissertation). University of Central Florida, Orlando, Florida.

Gorlach, M. (2000). Resultativeness: Constructions with phrasal verbs in focus. In E. Contini-Morava & Y. Tobin (Eds.), *Between grammar and lexicon* (pp. 255-287). John Benjamins: Amsterdam/Philadelphia.

Granger, S. & Meunier, F. (2008). Phraseology in language learning and teaching: Where to from here? In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 247-252). Amsterdam: John Benjamins.

Gries, St. Th., & Stefanowitsch, A. (2004). Co-varying collexemes in the into-causative. In M. Achard & S. Kemmer (Eds.), *Language, culture, and mind* (pp. 225-236). Stanford, CA: CSLI.

Grobe, C. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. *Research in the Teaching of English*, *15*, 75-85.

Hall, J. K., & Verplaetse, L. S. (2000). *Second and foreign language learning through classroom interaction*. Mahwah, NJ: Lawrence Erlbaum

Halliday, M.A.K. (1979). Differences between spoken and written language: Some implications for literacy teaching. In G. Page, J. Elkins & B. O'Connor (Eds.), *Communication through reading: Proceedings of the fourth Australian reading conference* (pp. 37–52). Adelaide, S.A.: Australian Reading Association.

Halliday, M. A. K. (1985). *Spoken and written language*. Oxford: Oxford University Press

Hancioğlu, N., Neufeld, S., & Eldridge, J. (2008). Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes*, *27*(4), 459-479.

Hansen, C. & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), A*cademic listening: Research perspectives* (pp. 241– 268). Cambridge: Cambridge University Press.

Harris, T. (2003). Listening with your eyes: The importance of speech-related gestures in the language classroom. *Foreign Language Annals, 36*(2), 180–187.

Hasan, R., (1984). Ways of saying: Ways of meaning. In R. Fawcett, M. A.K. Halliday, S. Lamb & A. Makkai (Eds.), *The semiotics of culture and language, language as social semiotic* (pp. 105–162). Frances Pinter: London.

Hincks, R. (2003). Pronouncing the Academic Word List: Features of L2 student oral presentations. In M. J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetics Sciences* (pp. 1545–1548). Barcelona, Spain. Retrieved from http://www.speech.kth.se/ctt/publications/papers03/icphs03_1545.pdf

Hinkel, E. (2006). Current perspectives on teaching the four skills. *TESOL Quarterly, 40*(1), 109–131.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8*(2), 689–696.

Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

Hood, S., & Forey, G. (2005). Introducing a conference paper: Getting interpersonal with your audience. *Journal of English for Academic Purposes*, *4*(4), 291-306.

Howell, P., & Au-Yeung, J. (1995). The association between stuttering, Brown's factors, and phonological categories in child stutterers ranging in age between 2 and 12 years. *Journal of Fluency Disorders*, *20*(4), 331-344.

Howell, P., & Au-Yeung, J. (2007). Phonetic complexity and stuttering in Spanish. *Clinical Linguistics & Phonetics*, *21*(2), 111-127.

Hsu, W. (2014). The most frequent opaque formulaic sequences in English-medium college textbooks. *System*, *47*, 146-161.

Hsueh-chao, M. H. & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 23*(1), 403–430.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of English the language.* Cambridge: Cambridge University Press.

Hulstijn, J. H., & Marchena, E. (1989). Avoidance: Grammatical or semantic causes? *Studies in Second Language Acquisition*, *11*(3), 241-255.

Hunston, S. (1994) 'Evaluation and organisation in a sample of written academic discourse'. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 191–218). London: Routledge

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S., Francis, G., & Manning, E. (1996). *Collins COBUILD grammar patterns 1: Verbs*. London: Harper-Collins Publishers.

Hutchinson, T. & Waters, A. (1987). *English for specific purposes: A learning-centered approach.* Cambridge: Cambridge University Press.

Hyland, K. (1994). Hedging in academic writing and EAF textbooks. *English for Specific Purposes*, *13*(3), 239-256.

Hyland, K. (2006). Disciplinary differences: Language variation in academic discourses. In K. Hyland & M. Bondi (Eds.), *Academic discourse across disciplines* (pp. 17–45). Frankfurt: Peter Lang

Hyland, K. (2009). Writing in the disciplines: Research evidence for specificity. *Taiwan International ESP Journal*, *1*(1), 5-22.

Hyland, K. (2013). *Discourse studies reader: Essential excerpts*. London: Bloomsbury Academic.

Hyland, K., & Shaw, P. (2016). Introduction. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (pp. 1–13). London: Routledge.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, *41*(2), 235-253.

Jones, J. (2007). Multiliteracies for academic purposes: Multimodality in textbook and computer-based learning materials in science at university. In A. McCabe, M. O'Donnel & R. Whittaker (Eds.), *Advances in language & education* (pp.103–121). London: Continuum.

Jones, S., & Sinclair, J. M. (1974). English lexical collocations. A study in computational linguistics. *Cahiers de Lexicologie, 24*, 15-61.

Jordan, R. R. (1997). *English for academic purposes: A guide and resource book for teachers*. Cambridge: Cambridge University Press.

Kelly, P. (1991). Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *International Review of Applied Linguistics in Language Teaching*, *29*(2), 135-149.

Kennedy, G. (2014). *An introduction to corpus linguistics*. London: Routledge.

Kiewra, K. A. (2002). How classroom teachers can help students learn and teach them how to learn. *Theory into Practice*, *41*(2), 71-80.

Kilby, D. (1984). *Descriptive syntax and the English verb*. Dover, England: Croom

Kim, Y. (2006). Effects of input elaboration on vocabulary acquisition through reading by Korean learners of English as a foreign language. *TESOL Quarterly*, *40*(2), 341-373.

Knodt, E. A. (2006). What is college writing for? In P. Sullivan & H. Tinberg (Eds.), *What is "college-level" writing*? (pp. 146-157). Urbana, IL: NCTE.

Kroll, B. (1979). A survey of the writing needs of foreign and American College Freshmen. *English Language Teaching Journal*, *33*(3), 219-27.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Clevedon, England: Multilingual Matters.

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal*, *75*(4), 440-448.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307-322.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 36-55.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15-30.

Lee, D. (2001). Defining core vocabulary and tracking its distribution across spoken and written genres: Evidence of a gradience of variation from the British National Corpus. *Journal of English Linguistics*, *29*(3): 250-278.

Lee, D. Y. (2006). Humour in spoken academic discourse. *Journal of Language, Culture and Communication, 8,* 49–68.

Lee, J. J. (2009). Size matters: An exploratory comparison of small-and large-class university lecture introductions. *English for Specific Purposes*, *28*(1), 42-57.

Leech, G. (1991). The state of the art in corpus linguistics. In B. Altenberg & K. Aijmer (Eds.), *English corpus linguistics* (pp. 8–29). London: Longman

Leech, G., Rayson, P. & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Longman.

Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes*, *22*, 42-53.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Hove, England: Language Teaching Publications.

Li, E. & Pemberton, R. (1994). An investigation of students' knowledge of academic and sub technical vocabulary. In *Joint seminar on corpus linguistics and lexicography* (pp. 183-196). Hong Kong: HKUST Language Center.

Liao, Y., & Fukuya, Y. J. (2004). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning*, *54*(2), 193-226.

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

Lin, C. Y. (2012). Modifiers in BASE and MICASE: A matter of academic cultures or lecturing styles? *English for Specific Purposes*, *31*(2), 117-126.

Lindquist, H. (2009). *Corpus linguistics and the description of English*. Edinburgh: Edinburgh University Press.

Lipka, L. (1972). *Semantic structures and word-formation: Verb-particle constructions in contemporary English*. Munich: Wilhelm-fink-Verlag.

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly*, *45*(4), 661-688.

Li, W., Zhang, X., Niu, C., Jiang, Y., & Srihari, R. (2003). *An expert lexicon approach to identifying English phrasal verbs*. In proceedings of the 41st annual meeting of the ACL (pp. 513- 520). Sapporo: Japan.

Long, M., & Richards, J. (1994). Series editors' preface. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. ix-x). Cambridge, England: Cambridge University Press

Lynch, T. (2011) 'Academic listening in the 21st century: Reviewing a decade of research', *Journal of English for Academic Purposes*, *10*(2), 79–88.

Makkai, A. (1972). *Idiom structure in English*. The Hague: Mouton.

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*(3), 183–198.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, *33*(3), 299-320.

McArthur, T. (1989). The long-neglected phrasal verb. *English Today*, *5*(2), 38-44.

McCarthy, M. (1990). *Vocabulary*. Oxford: Oxford University Press.

McCarthy, M., & O'Dell, F. (2004). *English phrasal verbs in use*. Cambridge: Cambridge University Press.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McIntyre, A. (2002). Idiosyncrasy in particle verbs. In N. R. Dehé, A. McIntyre & S. Urban (Eds.), *Verb-particle explorations* (95-118). Mouton de Gruyter: Berlin

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL: The University of Chicago Press.

Miller, L. (2002). Towards a model for lecturing in a second language. *Journal of English for Academic Purposes*, *1*, 145–162.

Miller, L. (2003). Developing listening skills with authentic materials. *ESL Magazine, 6*(2), 16-18.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Milton, J., Wade, J., & Hopkins, N. (2010) 'Aural word recognition and oral competence in a foreign language'. In R. Chacon-Beltran, C. Abello-Contesse, & M. Torreblanca- Lopez (Eds.), *Further insights into nonnative vocabulary teaching and learning* (pp. 83–97). Bristol: Multilingual Matters.

Mitchell, T. F. (1958). Syntagmatic relations in linguistic analysis. *Translations of the Philological Society*, *57*(1), 101-118.

Moon, R. (1997). Vocabulary connection: Multi-word items in English. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 40-63). Cambridge: Cambridge University Press.

Morell, T. (2000). *EFL content lectures: A discourse analysis of an interactive and a non-interactive style*. Spain: Departamento de Filologıa Inglesa, Universidad de Alicante (Working papers, 7).

Morita, N. (2004). 'Negotiating participation and identity in second language communities'. *TESOL Quarterly*, *38*(4), 573-603.

Mueller, G. (1980). Visual contextual cues and listening comprehension: An experiment. *Modern Language Journal*, *64*(3), 335-340.

Muñoz, V. L. (2015). The vocabulary of agriculture semi-popularization articles in English: A corpus-based study. *English for Specific Purposes*, *39*, 26-44.

Meyer .G. A (1975). *The two-word verb. A dictionary of the verb-preposition phrases in*

American English. The Hague & Paris: Mouton.

Nagy, W., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly*, *24*, 262–282. https://doi.org/10. 2307/747770

Nagy, W. & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly, 47*(1), 91 - 108.

Nassaji, H. (2006). The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy use and success. *The Modern Language Journal*, *90*(3), 387-401.

Nation, P. (1990) *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. UK: 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*. Boston, MA, USA: Heinle Cengage Learning

Nation, I. S. P. (2012). *The BNC/COCA word family lists*. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd edition). UK: Cambridge University Press.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam: Benjamins.

Nation, I. S. P., & Anthony, L. (2013). Mid-frequency readers. *Journal of Extensive Reading*, *1*, 5–16.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Nation, P., & Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, *23*(1), 35–41

Nation, P., & Waring, R. **(**1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy (*pp. 6–19*).* Cambridge, UK: Cambridge University Press.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle Cengage Learning.

Nattinger, J., & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford, UK: Oxford University Press.

Nelson, M. (2000). *A corpus-based study of business English and business English teaching materials (*Unpublished doctoral dissertation), University of Manchester, Manchester. Retrieved from http://users.utu.fi/micnel/thesis.html.

Nesi, H. (2012). Laughter in university lectures. *Journal of English for Academic Purposes, 11*(2), 79–89.

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics*, *11*(3), 283-304.

Neufeld, S., Hancioğlu, N., & Eldridge, J. (2011). Beware the range in RANGE, and the academic in AWL. *System*, *39*(4), 533–538.

O'Dowd, E. M. (1998). *Prepositions and particles in English: A discourse-functional account*. Oxford: Oxford University Press.

Okamura, A. (2009). Use of personal pronouns in two types of monologic academic speech. *The Economic Journal of Takasaki City University of Economics*, *52*(1), 17-26.

O'Keeffe, A., McCarthy, M. (2010). *The Routledge handbook of corpus linguistics*. London: Routledge.

O'Sullivan, M. (2006). Lesson observation and quality in primary education as contextual teaching and learning processes. *International Journal of Educational Development*, *26*(3), 246-260.

Palmer, F. R. (1968). *A linguistic study of the English verb*. Coral Gables, FL: University of Miami Press

Palmer, F. R. (1974). *The English verb*. London: Longman.

Paribakht, T. S., & Webb, S. (2016). The relationship between academic vocabulary coverage and scores on a standardized English proficiency test. *Journal of English for Academic Purposes*, *21*, 121-132.

Park, J. Y. (2012). A different kind of reading instruction: Using visualizing to bridge reading comprehension and critical literacy. *Journal of Adolescent & Adult Literacy*, *55*(7), 629-640.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-226). New York: Longman.

Pikulski, J.J. & Templeton, S. (2004). *Teaching and developing vocabulary: Key to long-term reading success*. Boston, MA: Houghton & Mifflin.

Powers, D. (1985). *A survey of academic demands related to listening skills*. (TOEFL Research Report 20). Princeton, NJ: Educational Testing Service.

Praninskas, J. (1972). *American university word list*. London: Longman.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review, 56*(2), 282–308.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513-536.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive grammar of the English language*. London and New York: Longman.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Richards, J. (1974). Wordlists: Problems and prospects. *RELC Journal, 5*(2)*,* 69-84.

Richards, J. (1990). *The language teaching matrix*. Cambridge, UK: Cambridge University Press.

Rodgers, M. P. H. (2013). *English language learning through viewing television: An investigation of comprehension, incidental vocabulary acquisition, lexical coverage, attitudes, and captions* (Unpublished doctoral dissertation). Victoria University of Wellington, Wellington, New Zealand.

Rodgers, M., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly*, *45*(4), 689-717.

Rodgers, M. & Webb, S. (2016). Listening to lectures. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (PP. 165-176). London: Routledge

Rosenhouse, J., Haik, L., & Kishon-Rabin, L. (2006). Speech perception in adverse listening conditions in Arabic-Hebrew bilinguals. *International Journal of Bilingualism, 10*(2), 119–135.

Rowley-Jolivet, E. (2002). Visual discourse in scientific conference papers. A genre-based study. *English for Specific Purposes, 21*(1), 19–40.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly*, *22*(1), 69 - 90.

Sawyer, J. H. (2000). Comments on Clayton M. Darwin and Loretta S. Gray's "Going after the phrasal verb: An alternative approach to classification". *TESOL Quarterly*, *34*(1), 151-159.

Schmitt, N. (2000). *Vocabulary in language teaching*. New York: Cambridge University Press.

Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing, and use*. Amsterdam: John Benjamins.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. UK: Palgrave Macmillan.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26-43.

Schmitt, N., & McCarthy, M. (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*(1), 17-36.

Schmidt-Reinhart, B. C. (1994). The effects of topic familiarity on second language listening comprehension. *The Modern Language Journal, 78*(2), 179-189.

Schmitt, N. & Schmitt, D. (2005). *Focus on vocabulary: Mastering the academic word list*. London: Longman.

Schmitt, N. & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4): 484-503.

Schmitt, N. Schmitt, D. & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55 - 88.

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, *36*(2), 145-171.

Sharpe, T. (2006). "Unpacking" scaffolding: Identifying discourse and multimodal strategies that support learning. *Language and Education, 20*(3), 211–231.

Shaw, P. (1991). Science research students' composing processes. *English for Specific Purposes*, *10*(3), 189-206.

Side, R. (1990). Phrasal verbs: Sorting them out. *English Language Teaching Journal, 44,* 144-152.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487-512.

Simpson-Vlach, R. C., & Leicher, S. (2006). *The MICASE handbook: A resource for users of the Michigan corpus of academic spoken English*. Ann Arbor: University of Michigan Press.

Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). *The Michigan corpus of academic spoken English*. Ann Arbor: The Regents of the University of Michigan. Retrieved October 5, 2017, from http://www.hti.umich.edu/m/micase/

Sinclair, J. (1991). *Corpus, concordance and collocation*. Oxford: Oxford University Press.

Siyanova, A., & Schmitt, N. (2007). Native and nonnative use of multi-word vs. one-word verbs. *IRAL-International Review of Applied Linguistics in Language Teaching*, *45*(2), 119-139.

Skyrme, G. (2010). 'Is this a stupid question? International undergraduate students seeking help from teachers during office hours'. *Journal of English for Academic Purposes*, *9*(3): 211–221.

Snow, C. (2010). Academic language and the challenge of reading for learning about science. *Science*, *328*(5977), 450 - 452.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31*(4), 577-607.

Stahl, S & Fairbanks, M. (1986). The effects of vocabulary instruction: A Model-based meta-analysis. *Review of Educational Research, 56*(1), 72-110.

Staehr Jensen, L. (2005). *Vocabulary knowledge and listening comprehension in English as a foreign language: An empirical study employing data elicited from Danish EFL learners*. (Unpublished doctoral dissertation). Copenhagen Business School, Copenhagen.

Sroka, K. A. (1972). *The syntax of English phrasal verbs*. The Hague and Paris: Mouton.

Stubbs, E. (2007). 'Here's one I prepared earlier': The work of scribe D on Oxford, corpus Christi College, MS 198. *The Review of English Studies*, *58*(234), 133-153.

Sueyoshi, A. & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*(4), 661–699.

Swales, J. (1990) *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press

Swales, J.M. (2001). 'Metatalk in American academic talk'. *Journal of English Linguistics*, *29*(1): 34–54.

Swales, J. (2004). *Research genres: Exploration and applications*. Cambridge: Cambridge University Press.

Taylor, C. (2008). What is corpus linguistics? What the data says. *ICAME Journal*, *32*, 179-200.

Thompson, S. (1994). Frameworks and contexts: A genre-based approach to analysing lecture introductions. *English for Specific Purposes, 13*(2), 171-186.

Thompson, P. (2005). Aspects of identification and position in intertextual reference in PhD theses. In E. Tognini-Bonelli & D. Camiciotti (Eds.), *Strategies in academic discourse* (pp. 31-50). Amsterdam: John Benjamins,

Thompson, P. (2006). A corpus perspective on the lexis of lectures, with a focus on economics lectures. In K. Hyland & M. Bondi (Eds.), *Academic discourse across disciplines* (pp. 253–270). New York, NY: Peter Lang.

Thompson, P. (2015). Changing the base for academic word lists. In P. Thompson, & G. Diani (Eds.), *English for academic purposes: Approaches and implications.* (pp. 317-341). Cambridge: Cambridge Scholars Publishing.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: Benjamins.

Tribble, C. (2002). Corpora and corpus analysis: New windows on academic writing. In J. Flowerdew (Ed.), *Academic discourse* (pp. 131–149). London: Longman.

Trimble, L. (1985). *English for science and technology*. Cambridge: Cambridge University Press.

Townsend, D., Filippini, A., Collins, P. & Biancarosa, G. (2012). Evidence for the importance of academic word knowledge for the academic achievement of diverse middle school students. *The Elementary School Journal*, *112*(3), 497 - 518.

Townsend, D. & Kiernan, D. (2015). Selecting academic vocabulary words worth teaching. *The Reading Teacher*, *69*(1), 113 - 118.

Valipouri, L., & Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes*, *12*(4), 248-263.

Van Zeeland, H., & Schmitt, N. (2012). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457-479.

Vidal, K. (2003). Academic listening: A source of vocabulary acquisition? *Applied Linguistics*, *24*(1), 56–89.

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, *61*(1), 219–258.

Vilkaitė, L. (2016). Formulaic language is not all the same: Comparing the frequency of idiomatic phrases, collocations, lexical bundles, and phrasal verbs. *Taikomoji Kalbotyra. lt*, (8)

Vongpumivitch, V., Huang, J. Y., & Chang, Y. C. (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.

Wang, J., Liang, S. L., & Ge, G. C. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, *27*(4), 442-458.

Ward, J. (2007). Collocation and technicality in EAP engineering. *Journal of English for Academic Purposes*, *6*(1), 18-35.

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, *28*(3), 170-182.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46-65.

Webb, S., & Chang, A. C.-S. (2012). Vocabulary learning through assisted and unassisted repeated reading. *The Canadian Modern Language Review, 68*(3), 267-290.

Webb, S., & Paribakht, T. S. (2015). What is the relationship between the lexical profile of test items and performance on a standardized English proficiency test? *English for Specific Purposes*, *38*, 34-43.

Webb, S., & Rodgers, M. (2009a). Vocabulary demands of television programs. *Language Learning*, *59*(2), 335-366.

Webb, S., & Rodgers, M. (2009b). The lexical coverage of movies. *Applied Linguistics*, *30*(3), 407-427

West, M. (1953). *A general service list of English words.* London: Longman.

White, E. M. (2007). *Assigning, responding, evaluating: A writing teacher's guide* (4th ed.). Boston, MA: Bedford/St. Martin's

Wilkins, D.A. (1972). *Linguistics in language teaching*. London: Arnold.

Willis, D. (2003). *Rules, patterns and words: Grammar and lexis in English language teaching*. Cambridge: Cambridge University Press.

Wood, D. (2007). Mastering the English formula: Fluency development of Japanese learners in a study abroad context. *JALT Journal, 29*(2), 209-230.

Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *The Canadian Journal of Applied Linguistics*, *12*(1), 39–57.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence, and classroom applications*. London/New York: Continuum.

Wood, D. (2015). *Fundamentals of formulaic language: An introduction*. London/New York: Bloomsbury.

Xue, G. & Nation, P. (1984). A university word list. *Language Learning and Communication*, *3*, 215–229.

Zhou, S. (2010). Comparing receptive and productive academic vocabulary knowledge of Chinese EFL learners. *Asian Social Science*, *6*(10), 14-19.

Appendix A

A list of the Lemmatized Phrasal Verbs on the PHaVE List

back up	come down	give out	move back	pull back	start out
backed up	came down	gave out	moves back	pulls back	started out
backing up	coming down	gives out	moved back	pulling back	starting out
backs up	comes down	giving out	moving back	pulled back	starts out
blow up	come off	given out	move in	pull out	step back
blows up	came off	give up	moves in	pulls out	stepping back
blew up	coming off	gave up	moved in	pulling out	stepped back
blown up	comes off	gives up	moving in	pulled out	steps back
blowing up	come on	giving up	move on	pull up	sum up
break off	came on	given up	moves on	pulls up	summed up
broke off	coming on	go ahead	moved on	pulling up	summing up
breaks off	comes on	goes ahead	moving on	pulled up	sums up
broken off	come out	went ahead	move out	put back	take back
breaking off	came out	gone ahead	moves out	putting back	takes back
break out	coming out	going ahead	moved out	puts back	taking back
broke out	comes out	go along	moving out	put down	took back
breaks out	come over	goes along	move up	putting down	taken back
broken out	came over	went along	moves up	puts down	take down
breaking out	coming over	gone along	moved up	put in	takes down
break up	comes over	going along	moving up	putting in	taking down
broke up	come through	go around	open up	puts in	took down
breaks up	came through	goes around	opens up	put off	taken down
broken up	coming through	went around	opening up	putting off	take in
breaking up	comes through	gone around	opened up	puts off	takes in
break down	come up	going around	pass on	put on	taking in
broke down	came up	go back	passed on	putting on	took in
breaks down	coming up	goes back	passing on	puts on	taken in
broken down	comes up	went back	passes on	put out	take off
breaking down	cut off	gone back	pay off	putting out	takes off
bring about	cutting off	going back	paid off	puts out	taking off
brings about	cuts off	go down	paying off	put up	took off
brought about	end up	goes down	pays off	putting up	taken off
bringing about	ends up	went down	pick out	puts up	take on
bring down	ending up	gone down	picks out	reach out	takes on
brings down	ended up	going down	picking out	reaches out	taking on
brought down	figure out	go in	picked out	reached out	took on

bringing down
bring back
brings back
brought back
bringing back
bring in
brings in
brought in
bringing in
bring out
brings out
brought out
bringing out
bring up
brings up
brought up
bringing up
build up
builds up
built up
building up
call out
called out
calling out
calls out
carry on
carries on
carrying on
carried on
carry out
carries out
carrying out
carried out
catch up
catches up
caught up
catching up
check out
checked out
checks out
checking out
clean up
cleans up

figures out
figuring out
figured out
fill in
filled in
fills in
filling in
fill out
filled out
fills out
filling out
find out
finds out
found out
finding out
follow up
follows up
following up
followed up
get back
got back
gets back
getting back
gotten back
get down
got down
gets down
getting down
gotten down
get in
got in
gets in
getting in
gotten in
get off
got off
gets off
getting off
gotten off
get out
got out
gets out
getting out

goes in
went in
gone in
going in
go off
goes off
went off
gone off
going off
go on
goes on
went on
gone on
going on
go out
goes out
went out
gone out
going out
keep on
keeps on
keeping on
kept on
keep up
keeps up
keeping up
kept up
lay down
laid down
laying down
lays down
lay out
laid out
laying out
lays out
line up
lines up
lining up
lined up
look around
looks around
looking around
looked around

pick up
picks up
picking up
picked up
point out
points out
pointing out
pointed out
go over
goes over
went over
gone over
going over
go through
goes through
went through
gone through
going through
go up
goes up
went up
gone up
going up
grow up
grows up
grew up
growing up
grown up
hand over
handed over
hands over
handing over
hang on
hung on
hanging on
hangs on
hang out
hung out
hanging out
hangs out
hang up
hung up
hanging up

reaching out
rule out
ruling out
ruled out
rules out
run out
runs out
running out
ran out
send out
sends out
sent out
sending out
set about
sets about
setting about
set down
sets down
setting down
set off
sets off
setting off
set out
sets out
setting out
set up
sets up
setting up
settle down
settles down
settled down
settling down
show up
showed up
showing up
shows up
shut down
shuts down
shutting down
shut up
shuts up
shutting up
sit back

taken on
take out
takes out
taking out
took out
taken out
take over
takes over
taking over
took over
taken over
take up
takes up
taking up
took up
taken up
throw out
threw out
throwing out
throws out
thrown out
turn out
turns out
turned out
turning out
turn around
turns around
turned around
turning around
turn back
turns back
turned back
turning back
turn down
turns down
turned down
turning down
turn off
turns off
turned off
turning off
turn over
turns over

cleaning up
cleaned up
close down
closed down
closing down
closes down
come about
came about
coming about
comes about
come along
came along
coming along
comes along
come around
came around
coming around
comes around
come back
came back
coming back
comes back
come in
came in
coming in
comes in

gotten out
get on
got on
gets on
getting on
gotten on
get through
got through
gets through
getting through
gotten through
get up
got up
gets up
getting up
gotten up
give back
gave back
gives back
giving back
given back
give in
gave in
gives in
giving in
given in

look back
looks back
looking back
looked back
look down
looks down
looking down
looked down
look out
looks out
looking out
looked out
look up
looks up
looking up
looked up
make up
makes up
making up
made up
make out
makes out
making out
made out

hangs up
hold back
holds back
held back
holding back
hold on
holds on
held on
holding on
hold out
holds out
held out
holding out
hold up
holds up
held up
holding up
hold on
holds on
held on
holding on
play out
plays out
played out
playing out

sits back
sitting back
sat back
sit down
sits down
sitting down
sat down
sit up
sits up
sitting up
sat up
slow down
slows down
slowed down
slowing down
sort out
sorts out
sorting out
sorted out
stand out
stood out
stands out
standing out
stand up
stood up
stands up
standing up

turned over
turning over
turn up
turns up
turned up
turning up
wake up
woke up
waking up
wakes up
woken up
walk out
walked out
walking out
walks out
wind up
winds up
winding up
wound up
work out
works out
worked out
working out
write down
writes down
writing down
wrote down
written down

Appendix B

A List of the 150 AWL Lemmatized Verbs

analyse	identify	affect	institute	comment	negate	hypothesise
analyze	identified	affected	instituted	commented	negated	hypotheses
analyzes	identifies	affecting	institutes	commenting	negates	hypothesised
analyzing	identifying	affects	instituting	comments	negating	hypothesising
analyzed	indicate	assist	institutionalise	compensate	philosophise	hypothesize
analyses	indicated	assisted	institutionalised	compensated	philosophises	hypothesized
analysed	indicates	assisting	institutionalises	compensates	philosophised	hypothesizes
analysing	indicating	assists	institutionalising	compensating	philosophising	hypothesizing
approach	interpret	categorise	institutionalized	consent	philosophize	implement
approached	interpreted	categorize	institutionalizes	consented	philosophizes	implemented
approaches	interpreting	categorised	institutionalizing	consenting	philosophized	implementing
approaching	interprets	categorises	invest	consents	philosophizing	implements
assess	involve	categorising	invested	constrain	publish	implicate
assesses	involved	categorized	investing	constrained	published	implicated
assessed	involves	categorizes	invests	constraining	publishes	implicates
assessing	involving	categorizing	itemise	constrains	publishing	implicating
assume	issue	compute	itemised	contribute	react	impose
assumes	issued	computed	itemises	contributed	reacted	imposed
assumed	issues	computerised	itemising	contributes	reacts	imposes
assuming	issuing	computing	maintain	contributing	reacting	imposing
benefit	labour	conclude	maintained	convene	register	integrate
benefits	labor	concluded	maintaining	convenes	registered	integrated
benefited	labored	concludes	maintains	convened	registering	integrates
benefiting	labors	concluding	normalise	convening	registers	integrating
concept	laboured	conduct	normalises	coordinate	rely	internalise
concepts	labouring	conducted	normalised	coordinated	relied	internalised
conceptualised	labours	conducting	normalising	coordinates	relies	internalises
conceptualising	vary	conducts	normalize	coordinating	relying	internalising
consist	varied	construct	normalizes	core	remove	internalize
consists	varies	constructed	normalizing	cores	removed	internalized
consisted	varying	constructing	normalized	coring	removes	internalizes
consisting	achieve	constructs	obtain	cored	removing	internalizing
constitute	achieved	consume	obtained	correspond	scheme	investigate
constitutes	achieves	consumed	obtaining	corresponded	schemed	investigated
constituted	achieving	consumes	obtains	corresponding	schemes	investigates
constituting	legislate	consuming	participate	corresponds	scheming	investigating
context	legislated	credit	participated	deduce	sequence	label
contexts	legislates	credited	participates	deduced	sequenced	labeled

contextualised	legislating	crediting	participating	deduces	sequences	labeling
contextualising	occur	credits	perceive	deducing	sequencing	labelled
contract	occurred	design	perceived	demonstrate	shift	labelling
contracts	occurring	designed	perceives	demonstrated	shifted	labels
contracted	occurs	designing	perceiving	demonstrates	shifting	occupy
contracting	proceed	designs	purchase	demonstrating	shifts	occupied
create	proceeded	equate	purchased	document	specify	occupies
created	proceeding	equated	purchases	documented	specified	occupying
creates	proceeds	equates	purchasing	documenting	specifies	parallel
creating	process	equating	range	documents	specifying	paralleled
define	processed	evaluate	ranged	dominate	validate	parallelled
defined	processes	evaluated	ranges	dominated	validates	parallelling
defines	processing	evaluates	ranging	dominates	validated	parallels
defining	require	evaluating	regulate	dominating	validating	phase
derive	required	feature	regulated	emphasise	access	phased
derived	requires	featured	regulates	emphasised	accessed	phases
derives	requiring	features	regulating	emphasising	accesses	phasing
deriving	research	featuring	reside	emphasize	accessing	predict
distribute	researched	finalise	resided	emphasized	approximate	predicted
distributed	researches	finalised	resides	emphasizes	approximated	predicting
distributing	researching	finalises	residing	emphasizing	approximates	predicts
distributes	respond	finalising	resource	ensure	approximating	project
establish	responded	finalize	resourced	ensured	attribute	projected
established	responding	finalized	resources	ensures	attributed	projecting
establishes	responds	finalizes	resourcing	ensuring	attributes	projects
establishing	section	finalizing	restrict	exclude	attributing	promote
estimate	sectioned	focus	restricted	excluded	code	promoted
estimated	sectioning	focused	restricting	excludes	coded	promotes
estimates	sections	focuses	restricts	excluding	codes	promoting
estimating	signify	focusing	secure	fund	coding	resolve
export	signified	focussed	secured	funded	commit	resolved
exported	signifies	focussing	secures	funding	commits	resolves
exporting	signifying	impact	securing	funds	committed	resolving
exports	source	impacted	seek	illustrate	committing	retain
finance	sourced	impacting	seeking	illustrated	communicate	retained
finances	sources	impacts	seeks	illustrates	communicated	retaining
financing	sourcing	injure	sought	illustrating	communicates	retains
financed	structure	injured	select	immigrate	communicating	stress
formulate	structured	injures	selected	immigrated	concentrate	stressed
formulated	structures	injuring	selecting	immigrates	concentrated	stresses
formulates	structuring	transfer	selects	immigrating	concentrates	stressing
formulating	acquire	transferred	survey	imply	concentrating	sum
function	acquired	transferring	surveyed	implied	confer	summed

functioned	acquires	transfers	surveying	implies	conferred	summing
functioning	acquiring	link	surveys	implying	conferring	sums
functions	layer	linked	locate	interact	confers	summarise
Justify	layered	linking	located	interacted	contrast	summarised
justified	layering	links	locating	interacting	contrasted	summarises
justifies	layers	domesticate	locates	interacts	contrasting	summarising
justifying	emerge	domesticated	adjust	maximise	contrasts	summarize
Grant	emerged	domesticating	adjusted	maximised	cycle	summarized
granted	emerges	domesticates	adjusting	maximises	cycled	summarizes
granting	emerging	undertake	adjusts	maximising	cycles	summarizing
Grants		undertaken	alter	maximize	cycling	
		undertakes	altered	maximized	debate	
		undertaking	altering	maximizes	debated	
		undertook	alters	maximizing	debates	
					debating	