Article:
Dang, TNY and Webb, S (2014) The lexical profile of academic spoken English. English for Specific Purposes, 33. pp. 66-76. ISSN 0889-4906

https://doi.org/10.1016/j.esp.2013.08.001

The lexical profile of academic spoken English

Abstract

This study investigated (a) the lexical demands of academic spoken English and (b) the coverage of the Academic Word List (AWL) in academic spoken English. The researchers analyzed the vocabulary in 160 lectures and 39 seminars from four disciplinary sub-corpora of the British Academic Spoken English (BASE) corpus: Arts and Humanities, Life and Medical Sciences, Physical Sciences and Social Sciences. The results showed that knowledge of the most frequent 4,000 word families plus proper nouns and marginal words provided 96.05% coverage, and knowledge of the most frequent 8,000 word families plus proper nouns and marginal words provided 98.00% coverage of academic spoken English. The vocabulary size necessary to reach 95% coverage of each sub-corpus ranged from 3,000 to 5,000 word families plus proper nouns and marginal words and 5,000 to 13,000 word families plus proper nouns and marginal words to reach 98% coverage. The AWL accounted for 4.41% coverage of academic spoken English. Its coverage in each sub-corpus ranged from 3.82% to 5.21%. With the help of the AWL, learners with knowledge of proper nouns and marginal words will need a vocabulary of 3,000 and 8,000 word families to reach 95% and 98% coverage of academic spoken English, respectively.

Key words: academic spoken English; text coverage; listening comprehension, the Academic Word List; corpus studies; vocabulary frequency

1. Introduction

Understanding academic spoken English such as lectures or seminars is one of the greatest challenges for second language (L2) learners at English-medium universities. A lack of vocabulary knowledge is one of the biggest reasons for these students' poor comprehension of academic spoken English (Kelly, 1991). Research has shown that vocabulary knowledge is a significant factor for successful listening comprehension (Stæhr, 2009). To help students improve their comprehension of academic spoken English, it is essential to explore the vocabulary size necessary to comprehend academic spoken English. Learning Coxhead's (2000) AWL might be the most effective way for L2 students to improve their comprehension of academic written text. However, it is not clear whether the AWL can improve comprehension of academic spoken text to the same degree that it improves comprehension of academic written text because there has been little research investigating this issue.

The aim of this study is to determine the coverage of the AWL in academic spoken English and the vocabulary size necessary to reach 95% and 98% coverage of academic spoken English both with and without the help of the AWL. By doing this, the present research may provide a vocabulary goal for English for Academic Purposes (EAP) courses which, when reached, may allow learners to understand academic spoken English. This study may also indicate the value of the AWL for improving comprehension of academic spoken English.

1.1. How many words do you need to know to comprehend academic spoken English?

One way to determine the lexical demands of text is to calculate the number of words needed to reach certain coverage points. Coverage is the percentage of known words in a text (Nation & Waring, 1997). It is useful to measure coverage because it may indicate the vocabulary size necessary for comprehension of text. Although there are many factors affecting comprehension, coverage may be the most influential factor (Laufer & Sim, 1985). There have been no studies investigating the coverage necessary for

comprehension of academic spoken English. However, L2 research on the coverage needed for comprehending written texts and general conversation may provide some indication of the vocabulary size needed for comprehension of academic spoken English.

Most L2 studies measuring the coverage necessary for comprehension have been conducted on written text. Laufer (1989) suggested that 95% coverage could lead to reasonable comprehension of an L2 academic text. However, Hu and Nation (2000) found that 98% coverage was needed for adequate unassisted reading comprehension of a relatively easy L2 fiction text. Schmitt, Jiang and Grabe (2011) found a linear relationship between lexical coverage and comprehension. Although they did not find a coverage figure that ensured comprehension, they suggested that the coverage level required may vary according to the degree of comprehension needed. They reported that 98% coverage may be necessary if comprehension test scores of 60% or higher are needed. This supports Laufer and Ravenhorst-Kalovski's (2010) suggestion that two lexical coverage thresholds based on the degree of comprehension are used: 95% for minimal and 98% for optimal comprehension.

While research findings on the relationship between coverage and reading comprehension have been consistent to some extent, studies investigating the relationship between coverage and listening comprehension have had rather inconsistent results. Bonk (2000) found that learners occasionally had good listening comprehension at 80-89% coverage and suggested that learners with effective coping strategies may achieve adequate listening comprehension at far below 95% coverage for short texts. However, further analysis of Bonk's results by Schmitt (2008) indicated that learners with coverage of 90% or less may not have had adequate listening comprehension while those with coverage of 95% or more had adequate comprehension. To date, Van-Zeeland and Schmitt's (2012) study may be the most comprehensive research on the relationship between lexical coverage and listening comprehension. Examining L1 and L2 learners' comprehension of informal narratives, they found that the lexical coverage necessary for listening comprehension depends on the desired degree of comprehension. They suggest that 98% may be a good coverage goal for "very high comprehension" while 95% may be the best

text coverage goal for "good but not necessarily complete" comprehension of informal narratives (p. 18-19).

The variation in findings suggests that the coverage necessary for comprehension may vary according to discourse type and the degree of desired comprehension. Comprehension of academic spoken English, on one hand, may be easier than comprehension of written texts or radio programs. This is because the aural input of academic spoken English is supported by speakers' facial expression or gestures (Harris, 2003) and other media such as handouts, textbooks and visual materials presented on the board or overhead projector (Flowerdew, 1994). On the other hand, comprehension of academic spoken English may be more difficult than comprehension of informal conversation (Van-Zeeland & Schmitt, 2012) because vocabulary used in informal conversation may consist of more high-frequency words than that used in academic spoken English.

Taken together, research suggests that coverage of 90%-99% may provide adequate comprehension of academic spoken English. The present study chose 95% and 98% coverage as the lower and upper boundaries indicating comprehension of academic spoken English. These coverage points were chosen because 95% and 98% coverage may indicate reasonable (Laufer, 1989) and ideal (Nation, 2006) comprehension of written text and these figures are supported by Laufer and Ravenhorst-Kalovski (2010) and Van-Zeeland and Schmitt (2012).

A considerable number of corpus-driven studies have provided information about the vocabulary size necessary to reach 95% and 98% coverage of different types of written discourse such as graded readers (Nation, 2006; Webb & Macalister, 2012), newspapers (Nation, 2006), children's literature (Webb & Macalister, 2012) and novels (Nation, 2006). However, fewer studies have paid attention to spoken discourse, and all of these studies have dealt with general conversation rather than academic spoken discourse. Nation (2006) found that including proper nouns, 3,000 word families accounted for 95% coverage and 6,000-7,000 words families provided 98% coverage of unscripted spoken

English. Similarly, 3,000 word families plus proper nouns and marginal words and 6,000-7,000 word families plus proper nouns and marginal words were needed to reach 95% and 98% coverage of TV programs (Webb & Rodgers, 2009a) and movies (Webb & Rodgers, 2009b). Van-Zeeland and Schmitt (2012) suggest that to reach 95% lexical coverage of spoken text, learners would need from 2,000 to 3,000 word families. Taken together, these studies suggest that coupled with proper nouns and marginal words, 2,000- 3,000 word families and 6,000-7,000 word families are needed to reach 95% and 98% coverage of general spoken English, respectively.

1.2. Coverage of the AWL in academic spoken English

Coxhead's (2000) AWL is the successor of Xue and Nation's (1994) University Word List. Based on the principle of specialized occurrence, range and frequency, the AWL lists 570 word families derived from a 3.5 million token corpus which consisted of four sub-corpora: arts, commerce, law and science. The AWL covered 10.0% of the tokens in Coxhead's academic corpus. The coverage provided by the AWL across the four disciplines ranged from 9.1% (science) to 12% (commerce).

Since the AWL was created, there have been a large number of studies investigating the distribution of the AWL in academic written English, most of which have reported positive results which are in line with Coxhead's (2000) findings. Cobb and Horst (2004) and Hyland and Tse (2007) are two studies examining the distribution of the AWL in multidisciplinary corpora. Cobb and Horst (2004) found that the AWL accounted for 11.6% coverage of their 14,283 token corpus of text segments in seven disciplines: linguistics, sociology, history, social psychology, development, medicine, and zoology from the *Learned* section of the Brown corpus. Hyland and Tse (2007) found that the AWL covered 10.6% of their 3.3 million token corpus of sciences, engineering, and social sciences, written by professional and student writers. Research on the coverage of the AWL in specific disciplines also supports Coxhead's (2000) findings. It has been shown that the AWL accounted for 10.07% coverage of medical research articles (Chen & Ge, 2007), 11.17% coverage of applied linguistics research papers (Vongpumivitch,

Huang & Chang, 2009), 9.06% coverage of agricultural research articles (Martínez, Beck & Panza, 2009), 11.3% coverage of engineering textbooks (Ward, 2009) and 10.46% coverage of the Hong Kong Financial Services Corpus (Li & Qian, 2010). The only exception is Konstantakis (2007) who reported rather low coverage of the AWL (4.66%). The low coverage of the AWL in this corpus may be because the corpus that was analyzed was made up of Business English course books.

Research investigating the AWL has demonstrated its value to comprehension of academic discourse. However, according to Nesi (2002), Thompson (2006) and Hyland and Tse (2007), findings are predominantly based on analysis of academic written text. Therefore, investigating whether the AWL can aid comprehension of academic spoken English is warranted. There have been few studies examining the distribution of the AWL in academic spoken English. The reason for this may be the difficulty in collecting and analyzing spoken data (Adolphs & Schmitt, 2003; Thompson, 2006). To date, there have been only three studies dealing with the coverage of the AWL in academic spoken text. Hincks (2003) found that the AWL accounted for only 2.4% of a 13,471 token collection of oral presentations done by learners of English. The poor coverage of the AWL in this corpus may be because the corpus that was examined was made up of academic speech produced by non native English speakers rather than native English speakers. Nesi (2002), in an attempt to develop an academic spoken word list to supplement the AWL based on the BASE corpus, found that her academic spoken word list consisted of words in the AWL and words not in the AWL. However, she did not report the number of word families in her academic spoken wordlist and the coverage of this wordlist. Neither did she mention how many words or what percentage of her word list overlapped with the AWL. To date, Thompson (2006) may provide the most comprehensive research on the coverage of the AWL in academic spoken English. As part of his research to create an academic lecture wordlist, Thompson compared the coverage of the AWL in academic lectures by analyzing the 160 lectures in the BASE corpus. The result shows that the AWL provided only 4.9 % coverage of the lectures.

Another question that remains to be answered is whether or not the AWL has an even distribution across disciplines of academic spoken English. In terms of academic written text, both Cobb and Horst (2004) and Hyland and Tse (2007) found that the AWL was not evenly distributed across disciplines. Cobb and Horst (2004) reported a variation in the coverage of the AWL across seven disciplinary sub-corpora with medicine having the lowest coverage (6.72%) and history the highest coverage (14.49%). Similarly, Hyland and Tse (2007) found an uneven distribution of the AWL across disciplines with the lowest coverage in sciences (9.3%) and the highest coverage in engineering (11.1%). In the field of academic spoken text, however, there has been no research examining the coverage of the AWL across disciplines. In fact, none of the three aforementioned studies of the coverage of the AWL in academic spoken English examined the coverage of the AWL in particular disciplines.

Last but not least, it should be noted that the AWL was made in relation to West's (1953) General Service List (GSL); that is, to be included in the AWL, a word family member must not belong to the most frequent 2,000 GSL word families. Although the GSL still works rather well (Nation & Hwang, 1995), it is quite old and does not include some current vocabulary (Nation & Webb, 2011). Nation's (2006) British National Corpus (BNC) lists, on the other hand, may better represent current vocabulary. In fact, a considerable number of the AWL word families are at the first, second and third 1,000 word levels of the BNC (Nation, 2004; Cobb, 2010). Therefore, it would be useful to determine how much the AWL actually helps learners who have already mastered the most frequent 1,000, 2,000 or 3,000 BNC word families comprehend academic spoken English. In other words, it may be important to examine what vocabulary size is needed to reach 95% and 98% coverage of academic spoken English if the AWL is known.

1.3. Research questions

The review has shown that numerous studies have investigated the vocabulary size necessary to reach 95% and 98% coverage of written text and spoken text in general communication. However, there is a need to explore the vocabulary size necessary to

reach 95% and 98% coverage of academic spoken English. Moreover, despite many studies examining the coverage of the AWL in written text, very few have investigated the coverage of this list in academic spoken text and none of these studies have examined the coverage of the AWL across disciplines. The present research will address these problems by answering the following five questions:

1. What vocabulary size is necessary to reach 95% and 98% coverage of academic spoken English?
2. What vocabulary size is necessary to reach 95% and 98% coverage of each sub-corpus presented in the BASE corpus?
3. What is the coverage of the AWL in academic spoken English?
4. What is the coverage of the AWL in each sub-corpus presented in the BASE corpus?
5. With the help of the AWL, what vocabulary size is needed to reach 95% and 98% coverage of academic spoken English?

2. Methodology

2.1. Materials

The BASE corpus, which consists of 160 lectures and 39 seminars recorded at the University of Warwick and the University of Reading between 2000 and 2005, was used in this study. This 1,691,997 token corpus was developed from four broad disciplinary sub-corpora: Arts and Humanities, Life and Medical Sciences, Physical Sciences and Social Sciences. Each sub-corpus includes 40 lectures and 10 seminars, except for Physical Sciences which only includes 9 seminars. The BASE corpus was chosen for two reasons. First, because it was developed from real university lecture and seminar discourse, the BASE corpus presents the academic spoken English that L2 learners often encounter when studying at English-medium universities. Second, the BASE corpus is the largest, academic, spoken British-English corpus with sub-corpora. Therefore, it appears to provide a useful comparison to the corpus analyzed in Coxhead's (2000)

study. The present study aims to compare the four disciplinary components of the BASE corpus, but not the lecture and seminar components. Lectures and seminars were analyzed together because there were too few tokens (441,841 tokens) in the seminar component to justify their separate analysis (Sinclair, 1991).

This study deals with receptive knowledge (listening comprehension). According to Nation and Webb (2011), the word family is the most suitable unit for counting in research focused on comprehension. The reason for this is that if learners know one or two members of the word family, little effort is needed for recognizing and understanding other family members (Nation, 2001; Nation & Webb, 2011). For example, if the word change is known, other members of its word family such as changing, changeable, and unchanged may be recognized and understood.

Text files of the transcript were used for the analysis. All words marked as inaudible in the transcript of the academic spoken corpus were removed because the present research only dealt with spoken language. There were 15,991 tokens (0.945%) presenting the speakers' names (e.g. *nf0157* or *sm0833*) and 2,041 tokens (0.121%) indicating the speakers' non-verbal actions such as *cough*, *sigh*, or *laugh* that were excluded from the analysis. Similarly, unfinished words (e.g. *wa-*, *ver-*) which accounted for 10,419 tokens (0.616%) of the whole corpus were also excluded. It should be stressed that the speakers' non-verbal actions and unfinished words, although not counted in the analysis, may contribute to the listeners' comprehension in spoken contexts (Harris, 2003). This feature demonstrates one difference between spoken and written discourse.

Phonetic transcriptions such as [k] or [ston], which accounted for 145 tokens (0.009%), were also removed from the corpus because they cannot be recognized by the RANGE program, and in the contexts of the lectures, they were likely to be known because they represented high-frequency words. Although these phonetic transcriptions may help the listeners recognize the mentioned words by modelling the way the speakers pronounced words, their tiny percentage in the corpus means that they would not have much effect on the results.

Contractions (e.g. *'cause*) and archaic spellings found in quotations (e.g. *beautifull*) were changed to match with spellings used in the BNC word lists. They accounted for 947 tokens (0.056%) and 137 tokens (0.008%) of the entire corpus, respectively. Without the changes in their spellings, these words would have been incorrectly categorized as being less frequent than the most frequent 14,000 word families. However, it should be noted that knowing the full forms of the words does not mean that the listeners can comprehend the words in their contracted forms. However, the small percentage of these changes suggests that they may have little impact on the results of the analysis.

Similarly, hyphens in most hyphenated items were replaced by spaces so that the words that made up hyphenated items would be classified according to their frequency in the BNC wordlists. For example, the hyphens in the words *full-time* and *part-time* were removed and they were then reclassified by the frequency of their single-word items. In contrast, hyphens in such words as *second-hand* and *peace-keeping* were removed and the items were joined to make the single words: *secondhand* and *peacekeeping*. The decision of whether to turn a hyphenated item into separate words or single words was made by checking whether its joined form appeared in the 14 baselists of the BNC or not. Moreover, the hyphenated items sometimes indicated that the speakers spelled the words letter by letter (e.g. *anarch A-N-A-R-C-H*, or *euhemerism E-U-H-E-M-E-R-I-S-M*). This accounted for 76 tokens (0.004%). The hyphens in these items were removed and spaces were inserted because this reflected exactly the way the listeners perceived the word by hearing the words spelt by the speakers. However, hyphens in such acronyms as B-B-C and O-D-A were deleted and the spaces were removed so that they appeared as their written forms BBC and ODA. This is because the way learners perceive these words in their spoken forms may be similar to that in their written form. Although the majority of hyphens in hyphenated items were removed, hyphens in formulas like *C-five-H-six* and *C-H-three-O-H* were kept. They accounted for 2,758 tokens (0.163%) of the corpus. This decision was made because if the hyphens were replaced by spaces, the formulas which represented low-frequency words *C-five-H-six* and *C-H-three-O-H* (e.g. C_5H_6 = cyclopentadiene, CH_3OH = methanol) would have become *C five H six* and *C H three O*

H which would then have been classified by the RANGE program as high-frequency words.

Although proper nouns are classified in the proper noun list (List 15), a number of proper nouns were incorrectly categorized by RANGE as *Not in the lists* (words that have lower frequency than the most frequent 14,000 word families). These items were reclassified and added to the proper noun list. Likewise, a certain number of marginal words such as *mm*, *mmhm*, *aagh* and *aahh* which accounted for 1,080 tokens (0.064%) of the corpus did not appear in the marginal word list (List 16) but were listed as *Not in the lists*. These items were reclassified and added to the marginal word list.

2.2. Analysis

The RANGE program (Nation & Heatley, 2002) was used to analyze the vocabulary in the BASE corpus. This computer program classifies vocabulary in a text according to whichever word lists are used with it. It can be downloaded from Paul Nation's website: http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx.

To find the vocabulary size necessary to reach 95% and 98% coverage of the corpus, Nation's (2006) 14 lists of word families from the 1,000 to 14,000 word levels were used with RANGE to show the 1,000 word level at which the words in the text appeared. These lists were created based on the range and frequency of occurrence of words in the BNC. Less frequent words which do not belong to the most frequent 14,000 word families were categorized by the RANGE program as proper nouns (List 15), marginal words (List 16), or *Not in the lists*. Proper nouns and marginal words (items which can only marginally be regarded as words (Nation & Webb, 2011) such as interjections, hesitation procedures, and exclamations [mm, mmhm, aagh and oh]) were included in the cumulative coverage at the 1,000 word level because EAP learners are likely to know or be able to recognize these words (Nation & Webb, 2011).

To determine the percentage of academic words in the BASE corpus and each sub-corpus, three baseword lists were used with RANGE. Baseword list 1 and 2 consist of the first and second 1,000 words of West's (1953) GSL and baseword list 3 is Coxhead's (2000) AWL. The analysis with RANGE and these lists provides the information about the coverage of each baseword list in the BASE corpus and each sub-corpus.

3. Results

Table 1 presents the cumulative coverage including proper nouns and marginal words for the BASE corpus and four sub-corpora. Coupled with proper nouns and marginal words, a vocabulary of 4,000 word families provided 96.05 % coverage and a vocabulary of 8,000 word families accounted for 98.00% coverage of the BASE corpus. The vocabulary necessary to reach 95% coverage differed between disciplines. Knowledge of the most frequent 3,000 word families plus proper nouns and marginal words was needed to reach 96.01% coverage of the Social Sciences sub-corpus. Knowledge of the most frequent 4,000 word families provided 96.16% and 96.03% coverage of the Arts and Humanities sub-corpus and the Physical Sciences sub-corpus, respectively. Knowledge of the most frequent 5,000 word families was necessary to reach 95.46% coverage of the Life and Medical Sciences sub-corpus. There were larger differences in the vocabulary necessary to reach 98% coverage between disciplines. The vocabulary size necessary to reach 98% coverage ranged from 5,000 to 13,000 word families plus proper nouns and marginal words. A vocabulary of the most frequent 5,000 word families plus proper nouns and marginal words was sufficient to reach 98.12% coverage of the Social Sciences sub-corpus. To reach 98% coverage of the other three sub-corpora, larger vocabulary sizes were need: 7,000 word families (Arts and Humanities), 10,000 word families (Physical Sciences) and 13,000 word families (Life and Medical Sciences). The results indicate that at both 95% and 98% coverage, Social Sciences was the least lexically demanding and Life and Medical Sciences was the most lexically demanding.

Table 1. Cumulative coverage including proper nouns and marginal words for the BASE corpus and each sub-corpus

Word list	BASE corpus	Arts and Humanities	Life and Medical Sciences	Physical Sciences	Social Sciences
1,000	87.54	87.68	85.59	87.72	89.14
2,000	92.94	92.97	91.16	92.97	94.59
3,000	94.70	94.81	93.19	94.72	96.01[a]
4,000	96.05[a]	96.16[a]	94.49	96.03[a]	97.44
5,000	96.83	96.95	95.46[a]	96.70	98.12[b]
6,000	97.35	97.61	96.05	97.10	98.54
7,000	97.68	98.01[b]	96.46	97.33	98.80
8.000	98.00[b]	98.36	96.84	97.64	99.03
9.000	98.25	98.58	97.15	97.90	99.23
10,000	98.44	98.73	97.46	98.14[b]	99.32
11,000	98.58	98.91	97.64	98.26	99.41
12,000	98.72	99.02	97.89	98.41	99.46
13,000	98.83	99.15	98.05[b]	98.48	99.52
14,000	98.97	99.23	98.36	98.59	99.57
Proper nouns	0.82	1.69	0.47	0.37	0.65
Marginal words	2.51	2.86	2.74	1.80	2.50
Not in the lists	1.03	0.76	1.66	1.40	0.43
Tokens	1,691,997	444,971	437,994	345,585	463,447

[a] Reaching 95% coverage

[b] Reaching 98% coverage

The distribution of the AWL in the BASE corpus and each sub-corpus is presented in Table 2. The AWL accounted for 4.41% coverage of the BASE corpus. This list was not evenly distributed across the four sub-corpora. It had the highest coverage in the Social Sciences sub-corpus (5.21%) and the lowest coverage in the Arts and Humanities sub-

corpus (3.82%). Coverage of the AWL in the Life and Medical Sciences sub-corpus and the Physical Sciences sub-corpus was 4.27% and 4.28%, respectively.

Table 2. Coverage of the BASE corpus and each sub-corpus by the General Service List (West, 1953) and the Academic Word List (Coxhead, 2000) (%)

Corpus	Proper noun	Marginal words	General Service List		AWL
			1st 1,000 words	2nd 1,000 words	
Arts and Humanities	1.69	2.86	81.54	3.46	3.82
Life and Medical Sciences	0.47	2.74	79.47	4.25	4.27
Physical Sciences	0.37	1.80	83.11	4.00	4.28
Social Sciences	0.65	2.50	82.84	3.59	5.21
BASE corpus	0.82	2.51	81.68	3.81	4.41

A considerable number of AWL word families appear in the first 3,000 word families of the BNC (Nation, 2004; Cobb, 2010). Consequently, to determine the vocabulary size necessary to reach 95% and 98% coverage of academic spoken English with the help of the AWL, the distribution of the AWL in the BNC word lists needed to be examined. Table 3 shows that 79 AWL word families (23,723 tokens) occurring in the BASE corpus were in the first 1,000 BNC word list. This accounted for 1.4020% coverage of the BASE corpus. 199 AWL word families (30,768 tokens) were in the second 1,000 BNC word list, which provided 1.8184% coverage of the BASE corpus. 87 AWL word families (7,005 tokens) were classified in the third 1,000 BNC word list, accounting for 0.4140% coverage of the BASE corpus. The number of AWL word families in the fourth and fifth 1,000 BNC word lists was 98 (7,677 tokens) and 62 (3,409 tokens). They provided coverage of 0.4537% and 0.2015%, respectively. By the sixth 1,000 word level, very few word families from the AWL appeared, and the coverage of these word families was less than 0.1%.

Table 3. The distribution of the AWL in the BNC lists for the BASE corpus

Word list	Tokens		Word families
	Raw	Percentage (%)	
1,000	23,723	1.4020	79
2,000	30,768	1.8184	199
3,000	7,005	0.4140	87
4,000	7,677	0.4537	98
5,000	3,409	0.2015	62
6,000	1,002	0.0592	19
7,000	559	0.0330	12
8.000	203	0.0120	7
9.000	196	0.0116	2
10,000	66	0.0040	2
11,000	10	0.0006	0
12,000	27	0.0016	0
13,000	2	0.0001	0
14,000	2	0.0001	0
Not in the lists	2	0.0001	1
Tokens	74,651	4.4119	568

Table 4 shows the cumulative coverage of the AWL items in each of the BNC lists in the second column. The third column of Table 4 presents the additional coverage of the AWL items that occur at a lower frequency level. The additional coverage was calculated by subtracting the total coverage of the AWL for the BASE corpus (4.4119%) from the cumulative coverage of the AWL items at each BNC word level. For example, for learners who know the most frequent 1,000 BNC word families and the AWL, their knowledge of the AWL would provide 3.0099% coverage of academic spoken text (4.4119%-1.402%).

Table 4. Support provided by the AWL for learners who know different amounts of vocabulary *as determined by the BNC word lists (%)*

Word list	Cumulative coverage of the AWL items in the BNC lists	Coverage of the remaining items in the AWL
1,000	1.4020	3.0099
2,000	3.2204	1.1915
3,000	3.6344	0.7775
4,000	4.0881	0.3238
5,000	4.2896	0.1223
6,000	4.3488	0.0631
7,000	4.3818	0.0301
8.000	4.3938	0.0181
9.000	4.4054	0.0065
10,000	4.4094	0.0025
11,000	4.4100	0.0019
12,000	4.4116	0.0003
13,000	4.4117	0.0002
14,000	4.4118	0.0001
Not in the lists	4.4119	0.0000

Table 5 illustrates the potential coverage with knowledge of the AWL for learners who know different amounts of vocabulary as determined by the BNC word lists. The potential coverage at a certain word level was the sum of the cumulative coverage including proper nouns and marginal words at that word level and the additional coverage of the AWL items that occur at a lower frequency level. For instance, learners may get potential coverage of 90.55% if they know the most frequent 1,000 BNC word families and the AWL (87.54% +3.0099%). Table 5 reveals that knowledge of the AWL can help learners who know the most frequent 3,000 word families in the BNC achieve 95.48% coverage of academic spoken English. Learners with the vocabulary size of 8,000 word families can reach 98.02% coverage if they know the AWL. It should be noted that at the 3,000 word level, coverage moved from a point at which learners may not be able to have

adequate comprehension (94.7%) to a point at which they may have adequate comprehension (95.48%) if they know the AWL.

Table 5. Potential coverage with knowledge of the AWL for learners who know different amounts of vocabulary *as determined by the BNC word lists*

Word list	Cumulative coverage including proper nouns and marginal words for the BASE corpus without knowledge of the AWL	Potential coverage with knowledge of the AWL
1,000	87.54	90.55
2,000	92.94	94.13
3,000	94.70	95.48[a]
4,000	96.05[a]	96.37
5,000	96.83	96.95
6,000	97.35	97.41
7,000	97.68	97.71
8.000	98.00[b]	98.02[b]
9.000	98.25	98.26
10,000	98.44	98.44
11,000	98.58	98.58
12,000	98.72	98.72
13,000	98.83	98.83
14,000	98.97	98.97
Proper nouns	0.82	
Marginal words	2.51	
Not in the lists	1.03	
Tokens	1,691,997	

[a] Reaching 95% coverage

[b] Reaching 98% coverage

4. Discussion

In answer to the first research question, a vocabulary size of 4,000 word families plus proper nouns and marginal words provides 95% coverage, and a vocabulary size of 8,000 word families plus proper nouns and marginal words provides 98% coverage of academic spoken English. Compared with the vocabulary size of 2,000-3,000 word families to reach 95% coverage and 6,000-7,000 word families to reach 98% coverage of general spoken English, the findings of the present study suggest that to achieve 95% and 98% coverage of academic spoken English, a larger vocabulary size is needed. In other words, learners will need knowledge of 1,000-2,000 more word families to reach 95% and 98% coverage of academic spoken English compared with general spoken English. This is in line with Adolphs and Schmitt's (2004) finding that learners needed a larger vocabulary to deal with academic/ training discourse than general conversation. It also suggests that the vocabulary for general spoken English is not sufficient for learners to be able to understand academic spoken English.

In answer to the second research question**,** the results indicated that there was great variation in the amount of vocabulary needed to reach 95% and 98% coverage of each discipline. With knowledge of proper nouns and marginal words, learners only need 3,000 word families to achieve 95% coverage and 5,000 word families to reach 98% coverage of the Social Sciences sub-corpus. In contrast, the vocabulary size necessary to reach 95% and 98% coverage of the Life and Medical Sciences sub-corpus was 5,000 word families plus proper nouns and marginal words and 13,000 word families plus proper nouns and marginal words, respectively. The variation between the vocabulary sizes necessary to reach 95% and 98% coverage of each sub-corpus supports Adolphs and Schmitt's (2004) finding that the amount of vocabulary needed for successful comprehension varies according to different types of spoken discourse.

The variation in vocabulary size needed to reach 95% and 98% coverage of academic spoken English in different disciplines suggests that different disciplines may have different lexical demands with some being more difficult to understand than others.

Although knowledge of technical vocabulary in one discipline may help L2 learners have better comprehension of academic spoken English in that discipline, the number of technical words, and the meanings of these items may change between disciplines. As a result, learners need to be aware that although they may have the vocabulary needed for adequate comprehension of one discipline, there may be lexical challenges to comprehension of other disciplines. At both 95% and 98% coverage, Social Sciences needed the smallest vocabulary sizes and Life and Medical Sciences needed the largest vocabulary sizes to reach those coverage points. This suggests that Social Sciences may be the least demanding discipline while Life and Medical Sciences is the most demanding discipline in terms of lexical coverage.

Receptive knowledge of the most frequent 4,000 rather than 8,000 word families should be aimed for as the minimum vocabulary size necessary to comprehend academic aural text for EAP learners for two reasons. First, in interactive communication, learners can make use of clues from gestures or use communicative strategies to facilitate their comprehension (Adolphs & Schmitt, 2003; Harris, 2003). This may help reduce the lexical burden in listening comprehension. Second, although 98% or higher may be ideal coverage, learners may still achieve adequate listening comprehension with coverage lower than 95% (Van-Zeeland & Schmitt, 2012). Hence, if the AWL is not known, then knowledge of the most frequent 4,000 word families may be the prerequisite vocabulary size in EAP courses. However, it should be noted that higher coverage should result in better comprehension.

In answer to the third research question, the AWL accounted for 4.41% of the tokens in the academic spoken corpus. This coverage is quite small compared with the coverage of the AWL in other studies of academic written corpora: 10.0% (Coxhead, 2000), 11.6% (Cobb & Horst, 2004), 10.6% (Hyland & Tse, 2007), 10.07% (Chen & Ge, 2007), 11.17% (Vongpumivitch et al., 2009), 9.06% (Martínez et al., 2009), 11.3% (Ward, 2009) and 10.46% (Li & Qian, 2010). However, the coverage provided by the AWL in this study is consistent with Thompson's (2006) findings. The coverage of the AWL found in Thompson (2006) was a bit higher (4.9%), perhaps because his corpus was limited to

lectures while the present research used data from both lectures and seminars. The modest coverage provided by the AWL in the academic spoken corpus may be because the AWL was developed from an analysis of written text. The large difference between the coverage provided by the AWL in spoken and written text suggests that the AWL may not fully cover academic vocabulary in academic spoken English.

In answer to the fourth research question, the AWL was not evenly distributed across disciplines. This is consistent with Cobb and Horst's (2004) and Hyland and Tse's (2007) findings. In the present study, the highest coverage of the AWL was in the Social Sciences sub-corpus and the lowest coverage was in the Arts and Humanities sub-corpus. This suggests that students planning to major in courses from the Social Sciences would benefit the most from learning this list while those whose major is within Arts and Humanities would get the least benefit. The reason for the higher coverage of the AWL in the Social Sciences sub-corpus, as Hyland and Tse (2007) suggest, may be the high frequency of words in the AWL that are common to business-oriented disciplines. The number of words related to business- oriented disciplines in the AWL may be the result of Coxhead's (2000) selection of disciplines. Her commerce sub-corpus consists of rather similar disciplines such as accounting, economics, and finance while other sub-corpora such as sciences include disciplines which share fewer similarities (e.g. geography, mathematics, and biology). As a result, in the present study, the AWL provided the greatest coverage in the Social Sciences sub-corpus which has business-oriented subjects.

Table 6. The distribution of the AWL in the BASE corpus in comparison with that in Coxhead (2000), Cobb and Horst (2004) and Hyland and Tse (2007) (%)

Rank in terms of coverage	Academic written corpus			Academic spoken corpus
	Coxhead (2000)	Cobb and Horst (2004)	Hyland and Tse (2007)	BASE corpus
1	12.0 (commerce)	14.49 (history)	11.1 (engineering)	5.21 (Social Sciences)
2	9.4 (law)	14.38 (social psychology)	11.0 (social sciences)	4.28 (Physical Sciences)
3	9.3 (arts)	13.44 (sociology)	9.3 (sciences)	4.27 (Life &Medical Sciences)
4	9.1 (science)	12.60 (linguistics)		3.82 (Arts & Humanities)
5		12.26 (development)		
6		7.31 (zoology)		
7		6.72 (Medicine(anatomy))		
Mean	9.95	11.6	10.47	4.40
SD	1.37	3.24	1.01	0.58

Interestingly, both the GSL and AWL provided a rather low cumulative coverage in the Life and Medical Sciences sub-corpus in comparison with other sub-corpora (see Table 2). One reason may be the large number of technical words appearing in this sub-corpus. This suggests that learners need another kind of vocabulary, namely technical words as well as high-frequency and academic words to understand academic spoken English in this field. This is supported by Chung and Nation's (2003) and Cobb and Host's (2004) studies. Chung and Nation (2003) found a fairly high percentage of technical vocabulary in their anatomy text (37.6%). Cobb and Horst (2004) found that the AWL provided the smallest coverage in their medicine sub-corpus in comparison with the other six disciplinary sub-corpora, which in their opinion, is the result of the high amount of specialized terminology in the medicine sub-corpus.

Although the AWL was not evenly distributed across sub-corpora of the spoken corpus, the difference in the coverage of the AWL between sub-corpora of the spoken corpus was smaller than the difference in the coverage of the AWL between sub-corpora of other written corpora. Table 6 shows that the means and standard deviations (SD) of the AWL in the BASE corpus are smaller than those in Coxhead (2000), Cobb and Horst (2004) and Hyland and Tse (2007). The small difference in the distribution of the AWL across each sub-corpus of the BASE corpus suggests that the AWL is still an effective tool to support listening to academic spoken English for different disciplines.

In answer to the fifth research question, with the help of the AWL, learners with a vocabulary size of 3,000 word families can reach 95% coverage of academic spoken English. To reach 98% coverage, a vocabulary size of 8,000 word families is needed. In contrast, if the AWL is not known, 4,000 and 8,000 word families are needed to reach 95% and 98% coverage, respectively.

Because the results showed that there were 79, 199 and 87 items from the AWL in the first three 1,000 word BNC lists, L2 learners who know the most frequent 3,000 BNC word families would only need to learn the remaining 205 word families from the AWL to reach 95% coverage of academic spoken English (see Table 3). As a result, the AWL

provides a smaller lexical burden for L2 learners to reach 95% coverage than the 1000 items at the fourth 1,000 word level. Therefore, although the AWL has lower coverage in academic spoken English than academic written English, it has value in helping learners save time and effort to reach 95% coverage.

Although it is not clear whether or not a new Academic Spoken Word List (ASWL) would provide a higher coverage than the AWL, the low coverage of the AWL in the BASE corpus suggests that research is warranted. Nesi (2002) and Thompson (2006) also suggest that it would be beneficial to create an ASWL to supplement the AWL. Moreover, because many word families in the AWL appear in the first 3,000 word families of Nation's (2006) BNC lists, it would be useful to create the ASWL within the BNC framework. This is in line with Cobb (2010) who suggests that a modified AWL should be developed within the BNC framework to reflect current vocabulary use. Although the GSL still has value, it is rather old and may not represent the high-frequency vocabulary used today (Nation & Webb, 2011). Therefore, it may be useful if the ASWL was developed based on recently developed wordlists such as Nation's (2006) BNC lists or Nation's (2012) BNC/COCA lists.

However, until a spoken academic word list is created, the AWL is still a valuable tool for supporting comprehension of academic spoken English for two reasons. First, the low variation in coverage of the AWL across disciplines in academic spoken English suggests that it can be used for EAP learners from different disciplinary backgrounds. Second, instead of learning 1,000 word families at the fourth 1,000 word level, with the help of the 570 item AWL, learners with a vocabulary size of 3,000 word families can reach 95% coverage of academic spoken English.

It should be noted that although vocabulary tends to be learned according to word frequency level (Schmitt, Schmitt, & Clapham, 2001), learning does not occur in 1,000 word units and some lower frequency words will be learned before mastery of higher frequency word levels (Webb & Chang, 2012). Thus, the cumulative coverage figures represent ideal rather than typical vocabulary development. The findings of this study

(and other studies examining the lexical profile of discourse types) provide support for a frequency based vocabulary learning program**.**

Several limitations of this study should be noted. First, this research is based on a British academic spoken text. To gain a complete picture about the issues in academic spoken English, similar research should be conducted in other varieties of academic spoken English. For example, it would be useful to examine the vocabulary necessary to reach 95% coverage of the Michigan Corpus of Academic Spoken English (MICASE) and the coverage of the AWL in this corpus. Second, using the RANGE program to analyze the data, this study is unavoidably affected by the limitations of the RANGE program mentioned by Nation and Webb (2011). RANGE is unable to distinguish between homographs (e.g. *kind* (generous) and *kind* (type)) and unable to count multiword items (e.g. *as well as*) as single items. It is also inconsistent in dealing with compound words. Moreover, RANGE treats an apostrophe as a word break and classifies some very low-frequency items as members of higher frequency word families. Third, although lexical coverage may be the most influential factor affecting comprehension (Laufer & Sim, 1985), it is important to note that there are other factors that may affect comprehension of academic spoken English such as L1 listening ability (Vandergrift, 2006), background knowledge and topic familiarity (Schmidt-Rinehart, 1994) or learners' strategic competence (Bonk, 2000). Fourth, this study does not compare the lexical demands of academic spoken English and the coverage of the AWL in academic spoken English of two discourse types (seminars and lectures) due to the small number of tokens in the seminars. However, it may be useful if future research investigates this issue because lexical demand may vary according to different types of spoken discourse (Adolphs & Schmitt, 2004).

5. Conclusion

This study has shown that to reach 95% coverage of academic spoken English, L2 learners need a vocabulary of the most frequent 4,000 word families plus proper nouns and marginal words. However, with knowledge of the AWL, learners can reach 95%

coverage with a vocabulary of the most frequent 3,000 word families plus proper nouns and marginal words. This research also revealed that although the AWL provided only 4.41% coverage of academic spoken English, it had a fairly low variation in coverage across disciplines in academic spoken English. As a result, the findings suggest that the AWL has value in supporting comprehension of academic spoken English.

Acknowledgements

The RANGE program and wordlists used in this study were downloaded from Paul Nation's website: http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx.

The transcriptions 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. Details of how to access to the corpus via the Oxford Text Archive can be found at http://www.coventry.ac.uk/base.

References

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

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

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

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

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

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: John Benjamins.

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

Flowerdew, J. (1994). Research of relevance to second language lecture comprehension-an overview. In J. Flowerdew (Ed.), *Academic listening: research perspectives* (pp. 7-29). 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.

Hincks, R. (2003, August). *Pronouncing the academic word list: features of L2 student oral presentations.* Paper presented at the Proceedings of the 15th International Congress of Phonetics Sciences, Barcelona, Spain. Retrieved from http://www.speech.kth.se/ctt/publications/papers03/icphs03_1545.pdf.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

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

Kelly, P. (1991). Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *IRAL, 29*(2), 135-149.

Konstantakis, N. (2007). Creating a business word list for teaching business English. *Elia, 7*, 79–102.

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, UK: Multilingual Matters.

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.

Laufer, B., & Sim, D. D. (1985). Measuring and explaining the reading threshold needed for English for academic purposes texts. *Foreign Language Annals, 18*(5), 405-411.

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

Martínez, I., Beck, S., & Panza, C. (2009). Academic vocabulary in agriculture research articles. *English for Specific Purposes, 28*, 183–198.

Nation, I. S. P. (2001). Learning vocabulay in another language. Cambridge: Cambridge University Press.

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 3-14). Amsterdam: John Benjamins.

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

Nation, I. S. P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts. Retrieved from http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx

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

Nation, I. S. 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: Cambridge University Press.

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

Nesi , H. (2002, August). *An English spoken academic wordlist*. Paper presented at the EURALEX 2002, Copenhagen, Denmark. Retrieved from http://www.euralex.org/elx_proceedings/Euralex2002/036_2002_V1_Hilary%20Nesi_An%20English%20Spoken%20Academic%20Wordlist.pdf

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

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

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

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

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

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.

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: Peter Lang.

Van-Zeeland, H., & Schmitt, N. (2012). Lexical coverage in L1 and L2 listening comprehension: the same or different from reading comprehension? *Applied Linguistics*, *early view*, 1-24. doi: 10.1093/applin/ams074

Vandergrift, L. (2006). Second language listening: listening ability or language proficiency? *Modern Language Journal, 90*, 6-18.

Vongpumivitch, V., Huang, J., & Chang, Y. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes, 28*, 33–41.

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

Webb, S.A. & Chang, A, C-S, (2012). Second language vocabulary growth. RELC Journal, 43(1), 113-126.

Webb, S., & Macalister, J. (2012). Is text written for children appropriate for L2 extensive reading? . *TESOL Quarterly, 47*(2), 300-322. doi: 10.1002/tesq.70

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

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

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

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