White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/151462/

Version: Accepted Version

Article:
Dang, TNY (2020) The Potential for Learning Specialized Vocabulary of University Lectures and Seminars Through Watching Discipline-Related TV Programs: Insights From Medical Corpora. TESOL Quarterly, 54 (2). pp. 436-459. ISSN: 0039-8322

https://doi.org/10.1002/tesq.552

© 2019 TESOL International Association. This is the post-peer reviewed version of the following article: Dang, TNY (2019) The potential for learning specialized vocabulary of university lectures and seminars through watching discipline-related TV programs: Insights from medical corpora. TESOL Quarterly: a journal for teachers of English to speakers of other languages and of standard English as a second dialect. which has been published in final form at https://doi.org/10.1002/tesq.552. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Reuse
See Attached

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Source: Dang, T, N. Y. (in press). The potential for learning specialized vocabulary of university lectures and seminars through watching discipline-related TV programs: Insight from medical corpora. Article accepted in *TESOL Quarterly* on 23 September 2019.

The potential for learning specialized vocabulary of university lectures and seminars through watching discipline-related TV programs: Insights from medical corpora.

Abstract

This study investigated the potential of discipline-related television programs as the sources for incidental learning of specialized vocabulary of university lectures and seminars. First, a Medical Spoken Word List (MSWL) of 895 specialized word-types was developed from a 556,074-words corpus of medical lectures and seminars based on a mixed method: corpus-driven analysis, specialized dictionary checking, and expert ratings. Then, an 11,036,771-word corpus of 37 medical television programs was developed and analyzed to examine the extent to which the MSWL words were encountered in these programs. Adopting 5 encounters or more, 10 encounters or more, 15 encounters or more, and 20 encounters or more as the frequency cut-off points at which incidental learning may happen, this study found that the number of MSWL words that met these cut-off points increased as the number of episodes, seasons, and programs increased. This indicated that discipline-related television programs are potential sources for incidental learning specialized vocabulary of lectures and seminars if these programs are watched regularly and in a sequential order.

Key words: corpus, incidental vocabulary learning, specialized vocabulary, medicine, television, frequency, repetition, word list

INTRODUCTION

Specialized vocabulary is essential for academic success at English-medium university programs, but it is frequently cited as one of the greatest challenges by second language (L2) learners studying in these programs (e.g., Evans & Morrison, 2011). Therefore, it is important for researchers and practitioners to help EAP/ESP learners to develop the knowledge of specialized vocabulary that they would encounter often in their academic studies. In response to this call,

researchers have created specialized word lists from corpora that represent the genres in which EAP/ESP learners would engage with in their studies. Most of these lists were based on the analysis of written materials. The very few lists investigating spoken discourse focused on the shared vocabulary of a range of academic disciplines. Creating lists that represent vocabulary in academic spoken discourse of a specific discipline is important because EAP/ESP students would need to understand not only reading materials but also lectures and seminars in their academic study (Dang, Coxhead, & Webb, 2017).

With respect to teaching and learning, to help learners study items from specialized wordlists, a principled vocabulary program should combine both deliberate learning and incidental learning (Nation, 2013; Schmitt, 2008; Webb & Nation, 2017). Deliberate learning means vocabulary is learned through tasks or exercises whose primary aim is to retain words in short and long term memory. Incidental learning means vocabulary is learned as a by-product of another task such as reading or listening texts. Although deliberate learning is essential for acquiring a large amount of specialized vocabulary, relying solely on deliberate learning is inefficient for several reasons. First, there are limits to how much vocabulary can be explicitly taught and learned in the classroom (Webb & Nation, 2017). Second, not all EAP/ESP teachers have sufficient background knowledge on learners' specific disciplines to effectively teach specialized vocabulary (Coxhead, 2018). Third, vocabulary development is an incremental process which requires many encounters for new words to be learned and knowledge of known words to be consolidated (Nation, 2013; Webb & Nation, 2017). Therefore, apart from deliberate learning, incidental learning of specialized vocabulary through being exposed to the target language outside the classroom is an invaluable supplementary resource for L2 vocabulary learning (Schmitt, 2008).

For incidental learning to happen, learners need to be exposed to a great deal of input. Unfortunately, in many EFL contexts, the amount of input, especially specialized spoken input, is very limited (Webb & Nation, 2017). Thus, it is crucial to identify potential resources for incidental learning of specialized vocabulary of university lectures and seminars. Both corpus-based studies (Csomay & Petrovíc, 2012; Rodgers & Webb, 2011; Webb & Rodgers, 2009) and intervention studies (Rodgers, 2013; Peters & Webb, 2018) have indicated that incidental

vocabulary learning can happen through viewing television programs, and therefore, television programs may be potential resources for incidental vocabulary learning in EFL contexts.

The extent to which television programs can help learners to learn specialized vocabulary of academic lectures and seminars, however, is less transparent. The common assumption is that television programs may be irrelevant resources to learn specialized vocabulary of university lectures and seminars due to the differences between the two genres. University lectures and seminars are likely to be more formal and academic (as opposed to the more informal and entertaining nature of television programs). However, if we consider this issue more carefully, discipline-related television programs may be potential resources for learning specialized vocabulary of university lectures and seminars. Previous research found that watching related television programs offers more opportunities for incidental vocabulary learning than watching unrelated programs (Rodgers & Webb, 2011; Webb, 2011; Webb, 2011). Additionally, specialized vocabulary tends to occur more frequently in specialized texts than non-specialized texts (Chung & Nation, 2003; Nation, Coxhead, Chung & Quero, 2016). It follows therefore that discipline-related programs may contain a considerable number of specialized words of university lectures and seminars. Watching these programs may then provide great opportunities for frequent encounters with these words and help incidental learning to happen. To date, no studies have investigated the potential of discipline-related television programs as sources for incidental learning of the specialized vocabulary that EAP/ESP students are likely to encounter in academic lectures and seminars.

To fill these gaps, the present study aims to (a) develop a list of specialized vocabulary of university lectures and seminars in medicine and (b) examine the potential for incidental learning of items in this list through watching medical television programs. It is important to investigate specialized vocabulary in academic speech of medicine. Dang and Webb (2014) found that academic speech from Life and Medical Sciences are more challenging in terms of vocabulary than those from Arts and Humanities, Physical Sciences, and Social Sciences; that is, to achieve reasonable comprehension of academic lectures and seminars, learners would need a larger vocabulary size in the case of Life and Medical Sciences (5,000 word-families) than in the case of other disciplines (3,000-4,000 word-families). However, no medical spoken wordlists are available while existing medical written wordlists may not be sufficient in helping learners deal

with academic speech because there might be differences between spoken and written language. Additionally, the growing number of English language medical television programs (e.g., *Grey's Anatomy, The Good Doctor*), which L2 learners can easily access through DVD, cable television, and online media websites, suggests that these programs may be potential resources for EAP/ESP students to learn specialized vocabulary of medical lectures and seminars incidentally.

Specialized vocabulary in medicine

There are two views towards defining specialized vocabulary (Liu & Lei, 2020; Nation, 2016). The narrow view defines vocabulary as those words that are common in a specific discipline or a group of disciplines but are uncommon in other disciplines or other groups of disciplines (e.g., Coxhead, 2000; Wang, Liang, & Ge, 2008). The broad view (e.g., Lei & Liu, 2016; Ha & Hyland, 2017; Lu, 2018) considers specialized words as those that are closely related to a particular discipline. They can range from items which are typically only known by specialists in that discipline (e.g., *aorta, renal*) to items also known by people who are not specialists in that discipline (e.g., *heart, blood*). This view takes into consideration the argument that many words may have high frequency in general use but also carry specialized meanings within a particular discipline, and they deserve to be classified as specialized vocabulary (e.g., Dang, Coxhead, & Webb, 2017; Gardner & Davies, 2014; Ha & Hyland, 2017; Lei & Liu, 2016; Lu, 2018). The broad view will be taken in the present study.

Specialized vocabulary is important because it can make up a large proportion of words in a specialized text. Let us take the field of medicine as an example. Specialized vocabulary accounts for 12.24%-31.75% of the words in medical texts (Chung & Nation, 2003; Hsu, 2013; Lei & Liu, 2016; Wang et al., 2008). This suggests that specialized vocabulary may present a great learning burden for L2 learners. In fact, specialized vocabulary was listed as one of the biggest challenges faced by L2 learners at English-medium universities (e.g., Evans & Morrison, 2011). Such situation highlights the need for EAP/ESP researchers and teachers to support L2 learners in the study of specialized vocabulary.

Much of the effort around specialized vocabulary in EAP/ESP research has focused on the development of wordlists from specialized corpora for L2 learners. Several corpus-based wordlists have been specifically created to serve the need of EAP/ESP students who wish to study medicine (Hsu, 2013; Lei & Liu, 2016; Wang et al., 2008). All of them are written

wordlists while available specialized spoken wordlists—Academic Spoken Word List (Dang et al., 2017), Soft Science Spoken Word List (Dang, 2018a), and Hard Science Spoken Word List (Dang, 2018b)—present *shared* words between medicine and other disciplines rather than *all* the words that occurred frequently in academic speech of medicine. Developing a medical spoken wordlist is important because vocabulary in spoken discourse may be different from that in written discourse. Moreover, previous research has suggested that there was a substantial variation in the lexical items of academic speech from different disciplines (Dang, 2018a), and the lexical demands of academic speech from Life and Medical Sciences are likely to be greater than those from other disciplines (Dang & Webb, 2014). The development of a medical spoken wordlist would shed further insights into the nature of vocabulary in medical spoken English as well as providing EAP/ESP students with a useful tool for vocabulary learning.

Television as a source for incidental learning of specialized vocabulary

Most research on incidental vocabulary learning has looked at learning from reading (e.g., Pellicer-Sánchez & Schmitt, 2010; Webb & Chang, 2015) and listening (e.g., van Zeeland & Schmitt, 2013; Vidal, 2003). In recent years, however, there has been increasing interest in vocabulary learning through audio visual input such as television programs. The motivation behind this trend may be the wide availability of English language television programs through DVDs, cable television, and online media websites. These programs are valuable sources of L2 spoken input in many EFL contexts, where there may be limited opportunities for L2 listening. Surveys with EFL learners revealed that watching L2 television programs is a more important source of out-of-class exposure to L2 than reading books (Peters, 2018). Experimental studies with EFL learners also indicated that L2 vocabulary may be learned incidentally through watching television programs (Nguyen & Boer, 2018; Peters et al., 2016; Peters & Webb, 2018; Rodgers, 2013).

Corpus-driven studies have examined incidental learning through watching television from two perspectives: (a) the number of words needed to comprehend television programs and (b) the frequency of re-occurrences of words in these programs. The first line of research draws on studies investigating the effect of lexical coverage on comprehension. Lexical coverage is the percentage of known words in a text (Nation & Waring, 1997). Because of its close relationship with comprehension (e.g., Schmitt, Jiang, &Grabe, 2011; van Zeeland & Schmitt, 2013), lexical

coverage is an important factor that allows us to determine the extent to which learners might be able to understand a text and incidentally learn vocabulary from that text. While the amount of lexical coverage needed for incidental learning may vary according to discourse types, it is commonly accepted that a coverage of 95% is necessary in the case of listening (van Zeeland & Schmitt, 2013). Previous research (Webb, 2011; Webb & Rodgers, 2009; Rodgers & Webb, 2011) consistently indicated that if programs were analyzed as a whole, 3,000 word-families plus proper nouns and marginal words would provide 95% coverage of television programs; however, there was a variation in the amount of vocabulary needed to reach 95% of each genre/program. This indicates that while 3,000 word-families is generally necessary for incidental vocabulary learning from television programs to happen, the vocabulary size required may vary from genre to genre and program to program.

The second line of research determines the extent to which incidental learning may occur through watching television programs by examining how often words reoccurred in these programs. As this line of research is directly related to the purpose of the current study, it will be discussed in more detail. This line of research builds upon empirical evidence that the more often words encountered in television programs, the more likely they are learned (Rodgers, 2013; Peters & Webb, 2018; Peters et al., 2016). Most previous studies have focused on incidental learning of low-frequency words. Webb and Rodgers (2009) analyzed vocabulary in a 264,384-word corpus made up of 88 television programs of various genres. They found that 69.15% of the low-frequency words in their corpus occurred only once or twice, and 15.6% were encountered 5 or more times. This indicated that incidental learning was unlikely to occur for most of low-frequency words with limited viewing over a variety of genres. However, Webb and Rodgers argued that the number of programs in their study was relatively small compared to the amount of time people watch television in their first language; therefore, if students watched television regularly over a long period of time, the potential for learning would increase. They also suggested that watching television programs from the same sub-genre that have similar topics and story lines may be an effective way to increase vocabulary learning through viewing.

Webb and Rodgers' (2009) suggestion was confirmed by subsequent studies. Rodgers and Webb (2011) compared the vocabulary in 142 episodes from six related television programs with those in 146 episodes from random television programs. They found that episodes from related

programs had lower vocabulary loads than episode from unrelated programs and that low-frequency words reoccurred more often in related programs than in unrelated programs. Webb (2011) further compared vocabulary in episodes from the same genres with those from different genres. Using the same corpus as Rodgers and Webb (2011), he categorized the six related television programs into three groups based on their genres: medical dramas, criminal forensic investigation dramas, and spy/action dramas. Webb also randomly grouped the 146 episodes from random television programs into three sets. He found that episodes within programs from the same genres had lower vocabulary load and higher percentage of low-frequency words reoccurrences than episodes from random programs. Together, Rodgers and Webb's (2011) and Webb's (2011) findings indicated that episodes from the same programs within the same genres may have greater potential for incidental vocabulary learning than episodes from unrelated programs. However, it should be noted that in Rodgers and Webb's (2011) and Webb's (2011) studies, each program only consisted of one season and each genre was represented by only two programs. Further research which focuses on a particular genre such as medical dramas and examines all seasons in the programs will provide further insight into the potential for vocabulary learning through television programs from the same genre.

The only study that has examined the potential for incidental learning of specialized vocabulary through watching discipline-related television programs was Csomay and Petrovíc's (2012) study. Defining specialized words as the words that appeared in discipline-related movies and television programs and had specialized meanings in a specialized dictionary, Csomay and Petrovíc (2012) created a specialized wordlist from a 128,897-word corpus of seven legal movies and a five-episode legal television program. Then, they examined the occurrences of these words in the same movie and television program corpus and found that words with 10 or more encounters accounted for 73.8%[1] of the specialized vocabulary in the corpus. Csomay and Petrovíc (2012) provide useful findings and highlight an area of incidental vocabulary learning that merits investigation. However, they did not intentionally focus on specialized vocabulary in academic lectures and seminars. Their specialized word list was developed from legal movies and television programs rather than from academic lectures and seminars. As a result, their study did not tell us the potential for learning the specialized words that EAP/ESP students would encounter in academic lectures and seminars in their future study.

Taken as a whole, the review of previous corpus-driven research on incidental learning through viewing has indicated that it is essential to investigate the occurrences of specialized vocabulary of academic lectures and seminars in discipline-related television programs, but no research has been conducted to address this need. If such research is conducted, it should be based on the analysis of vocabulary in a large corpus of academic lectures and seminars and a large corpus of multiple discipline-related television programs.

Number of encounters for incidental learning through viewing to happen

Incidental learning is an incremental process that needs a great amount of input (Webb & Nation, 2017). Experimental studies (Rodgers, 2013; Peters & Webb, 2018; Peter et al., 2016) found a relationship between the number of encounters and incidental vocabulary learning through viewing; that is, the more frequent words are encountered, the more likely they are to be learned. However, these studies did not indicate the frequency threshold at which incidental vocabulary learning through viewing happens. Consequently, previous corpus-driven research on viewing (Csomay & Petrovíc, 2012; Rodgers & Webb, 2011; Webb, 2011) set 10 or more times as the point at which incidental learning of new words happens, and 5-9 times as the points at which learners gain partial knowledge of known words. However, these cut-off points were based on studies with reading. Imagery presented in television can make learning words through viewing easier than through reading, but the on-line nature of viewing may make it more difficult to learn words through viewing than through reading (Rodgers, 2018). Therefore, it is unclear whether viewing requires more encounters for incidental learning than reading. In fact, Webb and Nation (2017) points out that there is no frequency threshold for incidental vocabulary learning to happen; instead, there is a relationship between the number of encounters and incidental learning. Thus, to provide better insights into the potential for incidental learning specialized vocabulary through viewing, rather than relying on a specific frequency cut-off point, the present study would use a range of cut-off points: (a) 5 or more encounters, (b) 10 or more encounters, (c) 15 or more encounters, and (d) 20 or more encounters. Words with encounters of 1-4 times are likely to offer a relative small amount of learning while words with higher numbers of encounters may have a greater likelihood of learning. Five encounters and 10 encounters were chosen because these cut-off points have been used by previous corpus-driven research on incidental vocabulary learning through viewing. The 15 encounter cut-off point was chosen because van Zeeland and Schmitt (2013) found that at least 15 encounters are needed for

incidental learning from listening. As students receive audiovisual support in the viewing condition, they may need fewer encounters in the viewing condition than in the listening condition. The 15 or more encounters, thus, is a useful cut-off point to examine the potential for incidental learning through viewing. The 20 encounter cut-off point was chosen because Uchihara, Webb, and Yanagisawa's (2019) meta-analysis of research on incidental learning revealed that the effect of frequency on incidental vocabulary learning is likely to remain prominent up to around 20 encounters.

The present study and research questions

The purpose of the present study is to examine the potential for learning specialized vocabulary through watching medical television programs. In particular, it aims to (a) develop a list of specialized vocabulary of university lectures and seminars in medicine and (b) examine the potential for incidental learning of items in this list through watching medical television programs. Unlike previous corpus-driven research on incidental vocabulary learning, this study would not make a list of specialized vocabulary from medical television programs but rather medical lectures and seminars. Also, instead of relying on one cut-off point, it would use a range of frequency cut-off points to examine the potential for incidental vocabulary learning. The study would shed light on the potential of discipline-related television programs for incidental learning specialized vocabulary of academic lectures and seminars. It seeks to address the following research questions:

1. What are specialized words in medical lectures and seminars?
2. To what extent can these words be encountered in medical television programs?

METHODOLOGY

Corpora

Two corpora were developed for this study. The medical academic spoken corpus (556,074 words) was created from transcripts of 32 university lectures and 17 university seminars in Health and Medical Sciences courses from five sources: the British Academic Spoken English corpus, Michigan Corpus of Spoken English, Pearson International Corpus of Academic English, Yale Open coursewares, and English as a Lingua Franca in Academic Setting corpus.

The medical television program corpus (11,036,771 words) was derived from transcripts of 2,073 episodes from 37 medical television programs. Following previous studies (e.g., Webb & Rodgers, 2009, Rodgers & Webb, 2011), these programs were selected based on the availability of scripts, genres, and popularity (See Appendix A for detailed information of these programs). A season refers to a short succession of episodes, lasting usually less than a year. A program consists of one or more seasons, which means that a program includes episodes from all seasons across time. Following previous research on the lexical demands of spoken discourse (e.g., Dang & Webb, 2014; Webb & Rodger, 2009), inaudible words such as stage commands, storyline (e.g., *country music playing, chuckles*) and speakers' name (e.g, *Chris, nf0157)* were removed from the transcripts. Only words that could be heard during the conversations were kept for the analysis. The two corpora developed in the present study are the largest medical spoken corpus and medical television program corpus that have ever been created.

Identifying specialized vocabulary in medical lectures and seminars

To identify specialized vocabulary in medical lectures and seminars, a mixed method was adopted: (a) corpus-driven analysis, (b) specialized dictionary checking, and (c) expert ratings. This follows the current trend in developing specialized wordlists (Dang, 2020; Liu & Lei, 2020; Nation, 2016). The corpus-driven analysis ensured that the initial list captures the most frequent, wide ranging, and unique lexical items in medical lectures and seminars. The specialized dictionary checking and expert ratings are essential. They took into account the fact that some general high-frequency words (e.g., *tissue, delivery*) also have specialized meaning and should be considered as specialized vocabulary and made sure that the list reflects the words that students are likely to meet in their discipline (Coxhead & Demecheleer, 2018).

In the corpus-driven analysis, word-type was chosen as the unit of counting of the Medical Spoken Word List (MSWL) because it is a common unit of counting of specialized wordlists (Lu, 2018; Nation, 2016; Liu & Lei, 2020). Tokens refer to the word forms occurring in a text (Nation, 2013). Repeated word forms are counted as separate tokens. In contrast, types are unrepeated word forms occuring in a text (Nation, 2013). For example, *counting words is difficult but it is fun* contains eight tokens but seven word-types because the word form *is* occurred twice. The selected items for the initial list should (a) be content words, (b) occur with relative frequency of at least 9.4 times per million in the medical spoken corpus, (c) appear in at

least 5 transcripts, and (d) have the keyness of 28.7 when comparing its frequency in medical speech (represented by the medical spoken corpus) than its frequency in general conversation (represented by Love, Dembry, Hardie, Brezina, and McEnery's (2017) Spoken BNC2014). Only content words were selected so that the MSWL includes meaningful items. The frequency and range criteria ensured that the list captures the most frequent and wide ranging words in medical lectures and seminars, while the keyness criterion ensured the specialized nature of the words; that is, the selected words have significant higher frequency in the medical speech than in general conversation. The frequency, range, and keyness cut-off points were set as the result of extensive experimentation which compared items included in or excluded from the MSWL at different cut-off points. These cut-off points were selected because unlike more lenient cut-off points, these cut-off points ensure that the MSWL consists of a relatively small number of items (fewer than 900 words); unlike stricter cut-off points, these cut-off points still allow learners to recognize a reasonable proportion of words in medical lectures and seminars (more than 13%). Heatley, Nation, and Coxhead's (2002) RANGE was used to analyze the frequency and range of items in the medical spoken corpus. This program lists the words that occurred in a text based on their frequency and range. Anthony's (n.d) Antconc was used to determine the keywords. This program compares the frequency of words in a specialized corpus and a reference corpus and generates a list of key words whose frequency in the specialized corpus is significantly higher than that in the reference corpus.

Items selected in the corpus-driven analysis were then checked in two well-known medical English dictionaries: the Merriam-Webster's medical English dictionary and Taber's Cyclopedic medical dictionary. These dictionaries were used by Lei and Liu (2016) to identify items for their medical written vocabulary list. Words that appeared in neither dictionary were removed.

The degree of technicality of items remaining after the specialized dictionary checking was then rated by two experts. The first expert had a BA degree in medicine and an MA and PhD degree in Applied Linguistics. The second expert had a BA degree in English language and 18-year experience working as a doctor. A semantic scale was used in the rating (Table 1). This scale was adapted from the scales used in previous research on developing specialized wordlists (Chung & Nation, 2003; Ha & Hyland, 2017; Lu, 2018). When the experts were not sure which points to give to a certain word, concordance lines of that word in the medical spoken corpus

were provided to help them make the decision. Words rated as 1 by both experts (e.g, *cent, fashion, chart*) were removed from the list.

[TABLE 1 NEAR HERE]

Analyzing vocabulary in medical television programs

To determine the extent to which the MSWL words are encountered in medical television programs, transcripts in the medical television program corpus were run through the RANGE program with the MSWL as the base word list. The occurrences of the MSWL words were examined from five aspects: (a) in episode 1 of season 1 of each program, (b) in season 1 of each program, (c) in each program, (d) in each group of programs that have the same lexical demand, and (e) in all 37 programs together. This method of analysis allowed us to systematically determine the potential for learning the MSWL words through watching a single episode, a single season, a complete program, a group of programs with the same lexical demand, and all programs together. The MSWL words were classified into five bands based on the number of encounters in the corpus: (a) 1-4 encounters, (b) 5 or more encounters, (c) 10 or more encounters, (d) 15 or more encounters, and (e) 20 or more encounters.

To determine the lexical demands of each program, Nation's (2012) BNC/COCA twenty-five 1,000 word-family lists were used with RANGE to show the 1,000-word levels (1,000-25,000) at which the word-families in the medical drama program occurred. The BNC/COCA lists are the largest and most recent and popular frequency-based wordlists of general English. Words which do not belong to the most frequent 25,000 word-families were classified by RANGE as proper nouns (list 31), marginal words (List 32), compounds (list 33), abbreviation (list 34), and *Not in the lists*. Proper nouns and marginal words that were listed by RANGE as *Not in the lists* were added to the relevant lists. Following previous research on lexical demands of movies and television programs (e.g., Webb & Rodgers, 2009), proper nouns (e.g., *Catherine, Justin*) and marginal words (e.g., *uhuh, hmhm*) were included in the cumulative coverage at the 1,000-word levels with the assumption that they have a low learning burden and are likely to be understood in context. In the present study, the lexical demands were represented by the number of word-families together with proper nouns and marginal words needed to reach 95% coverage of the

program. The 95% coverage figures is commonly accepted as the point at which L2 learners may achieve reasonable comprehension of spoken texts (van Zeeland & Schmitt, 2013).

RESULTS

Specialized vocabulary in medical lectures and seminars

A total of 895 word-types were selected for the MSWL (see Appendix B for MSWL). Table 2 demonstrates the distribution of MSWL words across general vocabulary represented by the BNC/COCA levels. The 1,000, 2,000, and 3,000 BNC/COCA word levels represent high-frequency words while those outside the most frequent 3,000 BNC/COCA word levels are mid and low-frequency words (Schmitt & Schmitt, 2014). Most of the MSWL words are general high-frequency words: 27.15% of the words appearing at the 1st 1,000-word level, 27.6% at the 2nd 1,000-word level, and 24.47% at the 3rd 1,000-word level. Words at lower 1,000-word levels accounted for 20.78% of the list.

[TABLE 2 NEAR HERE]

To check the validity of the MSWL, following previous research (e.g., Gardner & Davies, 2014; Lei & Liu, 2016), the coverage of the list across general spoken, academic spoken, and medical spoken corpora was calculated. Love et al.'s (2018) Spoken BNC2014 corpus (17,090,008 words) was used as the general spoken corpus. Dang et al.'s (2017) Academic Spoken Word List (ASWL) corpus (without the medical texts) (12,558,866 words) was used as the academic spoken corpus. The Spoken BNC2014 is the largest corpus which represents spontaneous spoken English while the ASWL corpus is the largest corpus which features academic spoken English. The MSWL covered 13.44% of the medical spoken corpus. This coverage is higher than the coverage in the academic spoken corpus (8.15%) and the general spoken corpus (3.27%). These findings suggest that the MSWL is a list of words that are used much more frequently in medical spoken English than in general academic spoken English and general spoken English.

Potential for incidental learning from medical television programs

Results of the lexical demand analysis indicated that as a whole, a vocabulary size of the most frequent 3,000 word-families are needed to achieve 95% coverage of medical television programs (see Table 1- Appendix C). However, the vocabulary sizes needed to reach 95%

coverage of each program varied. In fact, these programs can be classified into four groups: (a) four programs with the lexical demand of 2,000 word-families, (b) 18 programs with lexical demand of 3,000 word-families, (c) 14 programs with lexical demand of 4,000 word-families, and (d) one program with lexical demand of 5,000 word-families (see Appendices A and C for further information about the names and lexical demands of these programs).

Let us now look at the occurrence of the MSWL words in the medical television program corpus. An average of 130 out of 895 MSWL word-types (14.53%) occurred in a single episode. Most of the MSWL word-types that appeared in a single episode occurred less than 5 times (87.76%). In contrast, the percentages of word-types with more reoccurrences were much lower: 12.24% (5 or more encounters), 5.23% (10 or more encounters), 2.91% (15 or more encounters), and 0.90% (20 or more encounters). The results indicated that there is likely to be a very small number of MSWL word-types being learned through viewing a single episode.

However, the results of the analysis of vocabulary in a single season, a single program, each group of programs that have the same lexical demand, and all 37 programs together indicated that as the number of episodes increased, the percentage of MSWL word-types appearing rose from 14.53% (a single episode) to 48.16% (a single season), 65.36% (a single program), 85.48% (each group of programs that have the same lexical demand), and 100% (all programs together). A similar trend is seen with the number of times the word-types were encountered (see Table 3). The percentage of MSWL word-types encountered less than 5 times decreased from 87.76% (a single episode) to 66.50% (a single season), 50.78% (a single program), and 23.75% (each group of programs that have the same lexical demand). In contrast, the percentage of MSWL words encountered more than 5 times increased from 12.24% (a single episode) to 33.50% (a single season), 49.22% (a single program), and 76.26% (each group of programs that have the same lexical demand). Similar patterns were seen with those encountered more than 10 times, more than 15 times, and more than 20 times. The percentage of MSWL word-types encountered 10 times went up from 5.23% (a single episode) to 17.37% (a single season), 31.51% (a single program), and 66.63% (each group of programs that have the same lexical demand). The percentage of MSWL word-types countered 15 times rose from 2.91% (a single episode) to 11.05% (a single season), 22.99% (a single program), and 60.73% (each group of programs that have the same lexical demand). Likewise, there is an increase in the percentage of MSWL word-

types encounters 20 times from 0.90% (a single episode) to 8.18% (a single season), 18.61% (a single program) and 56.74% (a group of programs). When the vocabulary in all 37 programs was analyzed altogether, 99.44% of the MSWL word-types appeared at least 20 times, three word-types (*cellular, females, particles*) appeared 19 times, one word-type (*molecules*) appeared 6 times, and one word-type (*molecule*) appeared 7 times in these programs.

[TABLE 3 NEAR HERE]

DISCUSSION

The present study has expanded upon earlier research on specialized vocabulary and incidental vocabulary through watching television in two ways. First, it sheds light on the nature of specialized words in medical lectures and seminars. Second, it provides insight into the potential for incidental learning of these words in medical television programs.

To answer the first research question, 895 word-types occurred frequently in a range of medical lectures and seminars and had medical meanings. Nearly 80% of these words are among the most frequent 3,000 words, which are considered as general high-frequency words (Schmitt & Schmitt, 2014). Interestingly, these words are evenly distributed across the 1st, 2nd and 3rd 1,000-word levels. As the first study to explore specialized vocabulary in academic speech of a specific discipline, this study reinforces the claim that specialized vocabulary cuts through different layers of general vocabulary (e.g., Dang et al., 2017; Gardner & Davies, 2014; Lei & Liu, 2016; Nation, 2016), and provides further evidence for including general vocabulary with specialized meanings in specialized word lists (e.g, *cases, cell, tissue, vessel*).

To answer the second research question, there is likely to be very few MSWL words learned incidentally through watching a single episode because the majority of the MSWL words were encountered less than 5 times in a single episode. This suggests that watching a single episode from medical television programs will have very little value as an activity to incidentally learn specialized words of medical lectures and seminars. However, the results also indicate that regular viewing of medical television programs over a long period of time has a great potential for incidental learning of these words. As the number of episodes increased, the number of encounters of MSWL words in the programs increased significantly. The percentage of MSWL words encountered 10 or more times went up from 5% in a single episode to nearly 67% in a

single group of programs that have the same lexical demand. Similarly, the percentage of MSWL words encountered 15 or more times rose from 3% in a single episode to more than 60% in a single group of programs that have the same lexical demand, and the percentage of MSWL words encountered 20 or more times went up from less than 1% to nearly 57% in a single group of programs that have the same lexical demand. Importantly, nearly 100% of the MSWL words appeared at least 20 times when all 37 programs were considered together.

Although words with 10 or more encounters, 15 or more encounters, and 20 or more encounters may have a greater likelihood of learning than those with a smaller number of encounters, if we consider the cut-off point of 5 times or more, which was adopted by previous research (Csomay & Petrovíc, 2012; Rodgers & Webb, 2011; Webb, 2011) as the boundary where partial knowledge of known words is gained, there is a significant increase in the number of MSWL words encountered (from 12% in a single episode to more than 76% in a single group of programs that have the same lexical demand). This indicates that discipline-related programs may be potential sources for incidental learning of partial knowledge of known vocabulary. This finding is meaningful given the nature of specialized vocabulary. As found in this study and previous studies, a number of specialized words (e.g., *tissue*) have specialized meanings (e.g., a collection of cells which forms parts of humans, animals and plants) which are different from their meaning in everyday language use (e.g., very soft and thin paper). By seeing these words frequently in discipline-related television programs in different contexts (e.g., *soft tissue injuries, soft tissue damage, it's a disease that causes scar tissue*), learners may develop the awareness of their specialized meaning and use. Overall this study indicates that if medical television programs are regularly watched over a long period of time, there might be a potential for incidental learning of specialized vocabulary of medical lectures and seminars in terms of both breadth and depth. This is really meaningful as it suggests that simply watching discipline-specific television programs for entertainment, EAP/ESP learners may have opportunities to incidentally learn specialized vocabulary of academic lectures and seminars.

The present study provides an interesting approach towards investigating the potential for vocabulary learning through viewing television programs. While most previous corpus-driven research on incidental learning through viewing (e.g., Rodgers & Webb, 2011; Webb, 2011) focused on low-frequency words, this study focused on specialized vocabulary. The specialized

vocabulary used in the investigation of the potential for incidental learning through viewing was derived from a corpus of academic lectures and seminars rather than from a corpus of television programs. Moreover, instead of relying on a single frequency cut-off point, this study used a range of cut-off points to investigate the potential for incidental learning from viewing. As a result, it can provide solid evidence of the value of discipline-related television programs as a source for EAP/ESP students to learn the specialized vocabulary that they would likely encounter often in their target disciplines.

This study found that as a whole, a vocabulary size of the most frequent 3,000 word-families are needed to achieve 95% coverage of medical television programs; however the vocabulary sizes needed to reach 95% coverage of each program varied. A vocabulary size of the most frequent 3,000 word-families is needed to reach 95% coverage of nearly half of the programs in the corpus (see Tables 2 and 3 in the Appendix C). Yet four programs (*Casualty, Private Practice, the Clinic, The Doctor Blake Mysteries*) only required a vocabulary size of the most frequent 2,000 word-families (Table 4-Appendix C). In contrast, 14 programs required a vocabulary size of the most frequent 4,000 word-families (*A young doctor's notebook, Chicago Med, Children's Hospital, Combat Hospital, ER, House MD, Kingdom Hospital, MASH, Medical Investigation, Mental, Northern Exposure, Off the Map, The Good Doctor,* and *The Resident)* and one program (*Miami Medical*) even required a vocabulary size of 5,000 word-families to reach 95% coverage (Tables 5 & 6-Appendix C). The finding of the present study supports findings of previous studies that 3,000 word-families is needed to reach 95% coverage of television programs, but the lexical demands are likely to vary considerably between different programs (Rodgers & Webb, 2011; Webb, 2011; Webb & Rodgers, 2009).

PEDAGOGICAL IMPLICATIONS

While developing corpus-based specialized wordlists have received a great deal of interest from researchers, how to implement these lists in learning and teaching is an underexplored area of vocabulary research (Coxhead, 2018). The present study is among the very few that attempt to do so. To begin with, the MSWL developed in the present study captures specialized vocabulary in medical lectures and seminars and is a useful tool for EAP/ESP learners who plan to study or are already studying medicine at English-medium programs and EAP/ESP teachers who work

with these students for two reasons. First, scores in international standardized tests such as TOEFL and IELTS are usually used as the requirements for international students to study in English-medium university programs, but these tests do not measure any of the specialized nature of the fields that these students may enter. Second, while ESOL researchers and professionals have been aware of this problem, most of their efforts have focused on helping students deal with specialized vocabulary in written texts, less attention has been directed toward specialized vocabulary in lectures and seminars although these speech events are essential components of university study. By developing the MSWL from a corpus of medical university lectures and seminars, the present study contributes empirically based linguistic description of specialized vocabulary in spoken English. The list can be an add-on to existing written word lists which can inform the selection of words for classroom instruction, independent learning, and material development for EAP/ESP learners who plan to study or are already studying medicine at English-medium programs.

Not only identifying specialized words of medical lectures and seminars, this study also indicates that medical television programs may be potential resources for incidentally learning these words. This is meaningful given the limited input of specialized vocabulary in many EFL contexts. It is even meaningful when considering the fact that simply through regular watching of medical television programs, EAP/ESP students who plan to study medicine may learn the specialized vocabulary of medical lectures and seminars.

To optimize the opportunity for incidental learning the MSWL words through watching medical television programs, learners and teachers should considered these following principles. First, learners' motivation to learn through watching television programs is likely to depend on the extent to which they can understand the program (Webb, 2015). The present study shows that the lexical demands of medical television programs vary. Therefore, learners should watch programs that are below or relevant to their current vocabulary levels before moving on to programs that are beyond their level. That is, they should watch programs that require a vocabulary size of 2,000 word-families before moving on to those requiring a vocabulary size of 3,000 words, and then 4,000 words and 5,000 words. Sequencing the viewing in this way would create ideal conditions for incidental vocabulary learning to happen, because learners are exposed to authentic materials but still likely understand the programs. Learners can take Webb, Sasao, and

Balance's (2017) Updated Vocabulary Levels Test to determine their vocabulary levels and refer to Appendix A to choose programs that match their vocabulary levels. By providing the list of the lexical demands of each popular medical television program and categorizing them into groups, this study effectively responds to Webb's (2015) call for providing teachers and learners with lists of the lexical demands of each television program that can be potential used for extensive viewing.

Second, incidental learning should be combined with deliberate learning of the MSWL words because research has indicated the value of combining incidental learning and deliberate learning in vocabulary study (Nation, 2013; Schmitt, 2008). When viewing medical television programs, if there are words that learners are interested in, they are encouraged to check if these words appearing in the MSWL. If so, these items are worth their attention. Learners are encouraged to look up the meaning of these words in medical specialized dictionaries and examine their collocations by using concordance tools and transcripts of medical lectures and seminars from BASE/MICASE corpus. Raising learners' awareness of the specialized meaning and use of the MSWL words is particularly important because the present study found that a reasonable number of specialized vocabulary in medicine are high frequency in general conversation but also have specialized meaning and use.

Caution, however, should be taken when interpreting the finding of the present study. This study looked at the issue from the perspective of frequency while other factors such as cognates and prior vocabulary knowledge also contribute to incidental vocabulary learning through viewing (Peters & Webb, 2018). Several areas deserve attention of future research. First, intervention studies would provide further insight into the potential of medical television programs as sources for incidental learning specialized vocabulary. Second, Lin (2014) found that multi-words can be learned incidentally through watching television programs from a range of genres. It would be interesting to examine the potential for incidental learning of specialized multi-words through watching discipline-related television programs. Last but not least, it would be interesting to examine specialized vocabulary in other speech events such as labs and tutorials (Coxhead, Dang, & Mukai, 2017).

CONCLUSION

This study is the first attempt to identify specialized vocabulary in university lectures and seminars in medicine and to investigate the potential for incidental learning of these words from watching medical television programs. The 895-word MSWL developed in this study is in itself a useful instrument for EAP/EAP students who plan to study medicine in English-medium programs. The value of this study, however, is not just restricted within the area of medicine. By indicating that there is a great potential for incidental learning of specialized vocabulary of lectures and seminars though regular watching of medical television programs, the study suggests that discipline-related television programs may be potential resources for incidental learning of specialized vocabulary, and therefore, may serve as a bridge from entertainment to academic literacy for EAP/ESP learners. This is meaningful given the limited amount of specialized spoken input in many EFL contexts, and the fact that learners can learn specialized vocabulary of academic lectures and seminars simply through regularly viewing discipline-related television programs.

NOTES

[1] There are two possible reasons for this high percentage. First, Csomay and Petrovíc (2012) included high frequency words in their specialized word list if these words have specialized meanings. Second, their list was validated in the corpus from which it was developed.

ACKNOWLEDGMENTS

I am grateful to the Editor and the anonymous Reviewers for their constructive feedback, which has greatly improved this article. I would like to thank Averil Coxhead, Alice Deignan, and Mike Baynham for their comments on the early version of this paper and Cailing Lu and Linh Nguyen for acting as the expert raters of the word list. My thanks to the following publishers and researchers for their generosity in letting me use their materials to create the corpora: Cambridge University Press, Pearson, the researchers in the British Academic Spoken English corpus project, the Michigan Corpus of Spoken English, the English as a Lingua Franca in Academic Setting corpus project, the Spoken BNC2014 project, the Massachusetts Institute of Technology Open courseware project, the Stanford Engineering Open courseware project, the University of California, Berkeley Open courseware project, and the Yale University Open courseware project.

THE AUTHOR

Thi Ngoc Yen Dang is a Lecturer at the University of Leeds. She obtained her PhD from Victoria University of Wellington. Her research interests include vocabulary studies and corpus linguistics. Her articles have been published in *Language Learning, English for Specific Purposes,* and *Journal of English for Academic Purposes.*

REFERENCES

Anthony, L. (n.d.). *AntwordProfiler*. Retrieved from http://www.laurenceanthony.net/antwordprofiler_index.html

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins Publishing.http://dx.doi.org/10.1075/scl.23

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

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238.http://dx.doi.org/10.2307/3587951

Coxhead, A. (2018). Vocabulary and English for Specific Purposes Research: Quantitative and qualitative perspectives. London: Routledge.http://dx.doi.org/10.4324/9781315146478

Coxhead, A., Dang, T. N. Y., & Mukai, S. (2017). Single and multi-word unit vocabulary in university tutorials and laboratories: Evidence from corpora and textbooks. *Journal of English for Academic Purposes*, *30*, 66–78.http://dx.doi.org/10.1016/j.jeap.2017.11.001

Coxhead, A., & Demecheleer, M. (2018). Investigating the technical vocabulary of plumbing. *English for Specific Purposes*, *51*, 84–97.http://dx.doi.org/10.1016/j.esp.2018.03.006

Csomay, E., &Petrovíc, M. (2012). "Yes, your honor!": A corpus-based study of technical vocabulary in discipline-related movies and TV shows. *System*, *40*, 305–315.http://dx.doi.org/10.1016/j.system.2012.05.004

Dang, T. N. Y. (2018a). The nature of vocabulary in academic speech of hard and soft sciences, *English for Specific Purposes, 51*, 69-83.http://dx.doi.org/10.1016/j.esp.2018.03.004

Dang, T. N. Y. (2018b). The hard science spoken word list. *ITL – International Journal of Applied Linguistics*, *169*(1), 44–71.https://doi.org/10.1075/itl.00006.dan

Dang, T. N. Y. (2020). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 288-303). New York: Routledge.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning*, *67*(4), 959–997.https://doi.org/10.1111/lang.12253

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*, 66–76.http://dx.doi.org/10.1016/j.esp.2013.08.001

Evans, S., & Morrison, B. (2011). Meeting the challenges of English-medium higher education: The first-year experience in Hong Kong. *English for Specific Purposes*, *30*, 198–208.http://dx.doi.org/10.1016/j.esp.2011.01.001

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305–327.http://dx.doi.org/10.1093/applin/amt015

Ha, A. Y. H., & Hyland, K. (2017). What is technicality? A technicality analysis model for EAP vocabulary. *Journal of English for Academic Purposes*, *28*, 35–49.http://dx.doi.org/10.1016/j.jeap.2017.06.003

Heatley, A., Nation, I. S. P., & Coxhead, 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

Hsu, W. (2013). Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research*, *17*(4), 454–484.http://dx.doi.org/10.1177/1362168813494121

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

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.http://dx.doi.org/10.1016/j.jeap.2016.01.008

Lin, P. M. (2014). Investigating the validity of internet television as a resource for acquiring L2 formulaic sequences. *System*, *42*, 164-176.http://dx.doi.org/10.1016/j.system.2013.11.010

Liu, D. & Lei, L. (2020). Technical vocabulary. In S. Webb (ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 111-124). New York: Routledge.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), 319-344.

Lu. C. (2018). Investigating knowledge and use of technical vocabulary in Traditional Chinese Medicine (Unpublished PhD thesis). Victoria University of Wellington, Wellington, New Zealand.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.http://dx.doi.org/10.1017/CBO9781139858656

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins.http://dx.doi.org/10.1075/z.208

Nation, I. S.P, Coxhead, A., Chung, T. M., & Quero, B. (2016). Specialized word lists. In *Making and using word lists for language learning and testing* (pp. 146–151). Amsterdam: John Benjamins.http://dx.doi.org/10.1075/z.208

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.

Nguyen, C. D., & Boers, F. (2018). The effect of content retelling on vocabulary uptake from a TED talk. *TESOL Quarterly*. https://doi.org/https://doi.org/10.1002/tesq.441

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language*, *22*(1), 31–55.

Peters, E. (2018). The effects of out-of-class exposure to English language media on learners' vocabulary knowledge. *ITL-International Journal of Applied Linguistics*, *169*(1), 142–168.

Peters, E., Heynen, E., &Puimège, E. (2016). Learning vocabulary through audiovisual input: The differential effect of L1 subtitles and captions. *System*, *63*, 134–148. http://dx.doi.org/10.1016/j.system.2016.10.002

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*. https://doi.org/10.1017/S0272263117000407

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

Rodgers, M. P. H. (2018). The images in television programs and the potential for learning unknown words. *ITL-International Journal of Applied Linguistics*, *169*(1), 191-211.

Rodgers, M. P. H., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly*, *45*(4), 689–717. http://dx.doi.org/10.5054/tq.2011.268062

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363. http://dx.doi.org/10.1177/1362168808089921

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. http://dx.doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. http://dx.doi.org/10.1017/S0261444812000018

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). *The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies*. https://doi.org/10.1111/lang.12343

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479. http://dx.doi.org/10.1093/applin/ams074

Vidal, K. (2003). Academic listening: A source of vocabulary acquisition? *Applied Linguistics*, *24*(1), 56–89. http://dx.doi.org/10.1093/applin/24.1.56

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a Medical Academic Word List. *English for Specific Purposes*, *27*(4), 442–458. http://dx.doi.org/10.1016/j.esp.2008.05.003

Webb, S. (2011). Selecting television programs for language learning: Investigating television programs from the same genre. *International Journal of English Studies*, *11*(1), 117–135.

Webb, S. (2015). Extensive viewing: Language learning through watching television. In D. Nunan & J. C. Richards (Eds.), *Language learning beyond the classroom* (pp. 159–168). New York: Routledge. http://dx.doi.org/10.4324/9781315883472

Webb, S., & Chang, A. C.-S. (2015). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition*, *37*(4), 651–675. http://dx.doi.org/10.1017/S0272263114000606

Webb, S., & Nation, I. S. P. (2017). *How Vocabulary is Learned*. Oxford: Oxford University Press.

Webb, S., & Rodgers, M. P. H. (2009). Vocabulary demands of television programs. *Language Learning*, *59*(2), 335–366. http://dx.doi.org/10.1111/j.1467-9922.2009.00509.x

Webb, S., Sasao, Y., &Ballance, O. (2017). The updated Vocabulary Levels Test. *ITL – International Journal of Applied Linguistics*, *168*(1), 34–70. http://dx.doi.org/10.1075/itl.168.1.02web

(8,849 words)

TABLES

Table 1. Semantic scale used in the present study

Scale	Description
1	Word that has no relationship with medicine
2	Word that has a meaning related to medicine and is (almost) the same as the meaning in everyday language use.
3	Word that has a meaning related to medicine and is different from the meaning in everyday language use
4	Word that has only one (or more) meaning(s) and it is (they are) only related to medicine

Table 2. Distribution of the Medical Spoken Word List across the BNC/COCA levels

BNC/COCA levels	Number of word-types	Example
1,000	243	*See, blood, case*
2,000	247	*brain, risk, stroke*
3,000	219	*dose, cell, tissue*
Outside the most frequent 3,000 words	186	*transplant, membranes, urine*
Total	895	

Table 3. Encounters with the MSWL word-types in a single episode, a single season, and a single program

Number of encounters	Mean	SD
1-4 encounters		
A single episode	87.76	6.75
A single season	66.50	15.72
A single program	50.78	22.81
A group of programs with the same lexical demand	23.75	31.88
5 or more encounters		
A single episode	12.24	6.75
A single season	33.50	15.72
A single program	49.22	22.81
A group of programs with the same lexical demand	76.26	31.88
10 or more encounters		
A single episode	5.23	7.04
A single season	17.37	11.73
A single program	31.51	20.62
A group of programs with the same lexical demand	66.63	40.04
15 or more encounters		
A single episode	2.91	6.91
A single season	11.01	6.85
A single program	22.99	17.44
A group of programs with the same lexical demand	60.73	42.08
20 or more encounters		
A single episode	0.90	0.89
A single season	8.18	5.40
A single program	18.61	15.93
A group of programs with the same lexical demand	56.74	41.54

APPENDIX

Appendix A. Programs in the medical television program corpus

#	Name	Release time	Ranking in IMDB	Number of seasons	Mean of episodes per season	Varieties of English	Lexical demands (word families)
1	Casualty	1986	6.0	2	3	British	2,000
2	Private Practice	2007	6.6	6	18	American	2,000
3	The Clinic	2003	6.4	1	13	Irish	2,000
4	The Doctor Blake Mysteries	2013	8.1	2	8	Australian	2,000
5	3 Lbs	2006	6.9	1	15	American	3,000
6	A Gifted Man	2011	7.0	1	17	American	3,000
7	Body Of Proof	2010	7.1	3	8	American	3,000
8	Code Black	2015	8.1	3	15	American	3,000
9	Doc Martin	2004	6.0	4	9	American	3,000
10	Emily Owens MD	2012	7.6	1	9	American	3,000
11	Green Wing	2004	8.4	2	9	British	3,000
12	Grey's Anatomy	2005	7.6	15	21	American	3,000
13	Hawthorne	2009	6.2	2	9	American	3,000
14	Mercy	2009	7.0	1	22	American	3,000
15	NIP/TUCK	2003	7.7	7	14	American	3,000
16	Nurse Jackie	2009	7.7	7	11	American	3,000
17	Royal Pains	2009	7.1	8	12	American	3,000
18	Saving Hope	2012	7.6	5	17	Canadian	3,000
19	Scrubs	2001	8.4	9	20	American	3,000
20	The Knick	2014	8.5	2	10	American	3,000
21	The Mob Doctor	2012	6.5	1	6	American	3,000
22	The Night Shift	2014	7.3	3	12	American	3,000

23	A Young Doctor's Notebook	2012	7.9	1	4	British	4,000
24	Chicago Med	2015	7.5	4	17	American	4,000
25	Children’s Hospital	2010	7.8	3	7	American	4,000
26	Combat Hospital	2011	7.5	1	11	Canadian	4,000
27	ER	1994	7.7	15	20	American	4,000
28	House MD	2004	8.8	8	11	American	4,000
29	Kingdom Hospital	2004	6.8	1	13	American	4,000
30	MASH	1972	8.4	1	6	American	4,000
31	Medical Investigation	2004	7.1	1	2	American	4,000
32	Mental	2009	6.2	1	13	American	4,000
33	Northern Exposure	1990	8.3	6	18	American	4,000
34	Off The Map	2011	7.3	1	13	American	4,000
35	The Good Doctor	2017	8.4	2	13	American	4,000
36	The Resident	2018	7.0	2	11	American	4,000
37	Miami Medical	2010	7.3	1	6	American	5,000

Appendix B. Medical Spoken Word List

Level 1: MSWL words at the 1st 1,000 BNC/COCA word levels

Sub-list 1

1	get	11	system	21	health	31	basically	41	problems
2	blood	12	called	22	part	32	order	42	force
3	see	13	come	23	water	33	rate	43	general
4	different	14	take	24	bone	34	left	44	control
5	time	15	little	25	move	35	cause	45	causes
6	actually	16	problem	26	able	36	drug	46	terms
7	same	17	give	27	side	37	try	47	history
8	heart	18	point	28	group	38	gets	48	comes
9	body	19	normal	29	case	39	certain	49	area
10	course	20	high	30	form	40	pain	50	level

Sub-list 2

51	show	61	present	71	minute	81	takes	91	line
52	care	62	idea	72	experience	82	red	92	higher
53	drugs	63	simple	73	groups	83	education	93	levels
54	action	64	shown	74	areas	84	weight	94	cases
55	treatment	65	involved	75	moving	85	children	95	situation
56	learning	66	sense	76	given	86	shows	96	rest
57	difference	67	drop	77	human	87	answer	97	set
58	skin	68	test	78	hospital	88	imagine	98	doctors
59	taking	69	amount	79	certainly	89	learn	99	stage
60	doctor	70	based	80	open	90	hearing	100	growth

Sub-list 3

101	early	111	presentation	121	expect	131	beginning	141	tests
102	systems	112	lose	122	related	132	field	142	die
103	space	113	leg	123	allows	133	stone	143	issue
104	natural	114	basic	124	closed	134	types	144	position
105	treat	115	air	125	clearly	135	carry	145	collecting
106	allow	116	clear	126	fast	136	base	146	neck
107	light	117	wall	127	parts	137	grow	147	differences
108	quickly	118	machine	128	showed	138	lead	148	caused
109	forms	119	major	129	movement	139	local	149	term
110	issues	120	ability	130	state	140	legs	150	secondary

Sub-list 4

151	aware	161	eye	171	involves	181	causing	191	signs
152	subject	162	degree	172	learned	182	complete	192	waste
153	naturally	163	leads	173	formed	183	management	193	act
154	add	164	educational	174	power	184	self	194	figure
155	bones	165	relationship	175	support	185	deep	195	ear
156	count	166	draw	176	ball	186	faster	196	healthy
157	expressed	167	hospitals	177	death	187	setting	197	rates
158	expression	168	moves	178	lift	188	appear	198	animal
159	services	169	view	179	smoking	189	step	199	bodies
160	shape	170	follow	180	walls	190	forces	200	central

Sub-list 5

201	courses	211	painful	221	experienced	231	continuing	241	experiencing
202	express	212	addition	222	fields	232	involve	242	sexually
203	soft	213	appears	223	treatments	233	simpler	243	untreated
204	controlled	214	amounts	224	deaths	234	supporting		
205	additional	215	dr	225	adding	235	admitted		
206	drops	216	concern	226	acts	236	appearance		
207	stages	217	markers	227	poorly	237	findings		
208	wave	218	exact	228	involvement	238	protective		
209	breath	219	unusual	229	movements	239	specialty		
210	experiences	220	birth	230	specialist	240	suggestions		

Level 2: MSWL words at the 1st 1,000 BNC/COCA word levels

Sub-list 1

1	pressure	11	common	21	increase	31	chest	41	alcohol
2	disease	12	resistance	22	slide	32	active	42	range
3	flow	13	specific	23	population	33	available	43	increased
4	process	14	diseases	24	region	34	mass	44	constant
5	medical	15	muscle	25	development	35	model	45	physical
6	measure	16	medicine	26	ray	36	properties	46	unit
7	concentration	17	practice	27	detail	37	conditions	47	damage
8	risk	18	develop	28	image	38	culture	48	loss
9	brain	19	rays	29	effect	39	section	49	energy
10	surface	20	evidence	30	developed	40	role	50	examination

Sub-list 2

51	imaging	61	attitude	71	department	81	adult	91	technology
52	lower	62	condition	72	developing	82	community	92	complications
53	positive	63	iron	73	site	83	nervous	93	exam
54	associated	64	units	74	apply	84	professional	94	features
55	channels	65	access	75	bleeding	85	advantage	95	speech
56	direction	66	current	76	affect	86	detect	96	background
57	environment	67	images	77	average	87	slides	97	recognize
58	regions	68	describe	78	production	88	basis	98	balance
59	social	69	design	79	reaction	89	exposed	99	pattern
60	complicated	70	directly	80	approach	90	increases	100	reduce

Sub-list 3

101	standard	111	project	121	breast	131	female	141	blocks
102	attention	112	typical	122	compared	132	mouse	142	commonly
103	identify	113	trial	123	delivery	133	nerve	143	designed
104	muscles	114	bind	124	diet	134	knee	144	measures
105	quality	115	memory	125	grade	135	pump	145	refer
106	contact	116	combination	126	separate	136	stretch	146	selection
107	details	117	series	127	male	137	compare	147	supply
108	stress	118	exposure	128	scale	138	correct	148	attached
109	deliver	119	fail	129	speed	139	models	149	identified
110	operation	120	increasing	130	affects	140	application	150	indicate

Sub-list 4

151	operate	161	lab	171	physically	181	referred	191	reactions
152	products	162	activities	172	adults	182	species	192	stroke
153	alive	163	assume	173	exchange	183	breathing	193	affected
154	injury	164	benefit	174	generation	184	channel	194	applied
155	practices	165	bound	175	laboratory	185	illness	195	benefits
156	prevent	166	cultures	176	pressures	186	bleed	196	expose
157	reduced	167	maximum	177	circumstances	187	concentrate	197	failing
158	reliable	168	processing	178	damaged	188	established	198	stable
159	sheet	169	product	179	events	189	mice	199	performance
160	wound	170	progress	180	females	190	motor	200	pregnancy

Sub-list 5

201	repair	211	agent	221	resistant	231	reduction	241	attend
202	states	212	aids	222	risks	232	select	242	complication
203	upper	213	operations	223	actively	233	combined	243	progressive
204	divide	214	attempt	224	alcoholic	234	immediate	244	examined
205	indicated	215	directed	225	gain	235	delivering	245	medically
206	operating	216	smooth	226	indication	236	examiner	246	examining
207	sites	217	affecting	227	ace	237	indicates	247	labs
208	agents	218	fold	228	attach	238	environmental		
209	relax	219	intensive	229	examine	239	gather		
210	transfer	220	reactive	230	extend	240	injuries		

Level 3: MSWL words at the 1st 1,000 BNC/COCA word levels

Sub-list 1

1	cancer	11	failure	21	structure	31	surgery	41	infection
2	cells	12	gene	22	response	32	joint	42	factor
3	cell	13	proteins	23	virus	33	phase	43	consultation
4	patients	14	clinical	24	assessment	34	lecture	44	carbon
5	patient	15	communication	25	genes	35	significant	45	device
6	DNA	16	potential	26	data	36	stem	46	methods
7	protein	17	vessels	27	tissues	37	complex	47	organ
8	tissue	18	hip	28	vessel	38	tube	48	sequence
9	molecules	19	volume	29	factors	39	primary	49	functions
10	function	20	cancers	30	molecule	40	biological	50	genetic

Sub-list 2

51	dose	61	overall	71	output	81	variety	91	initial
52	technique	62	behaviour	72	define	82	solution	92	relative
53	focus	63	analysis	73	structures	83	transport	93	internal
54	negative	64	generate	74	definition	84	external	94	behave
55	frequency	65	experiment	75	filter	85	host	95	infectious
56	evaluation	66	mechanism	76	mild	86	theory	96	molecular
57	respond	67	reflection	77	bacteria	87	communicate	97	mortality
58	assess	68	symptoms	78	circulation	88	linked	98	sensitive
59	concept	69	therapy	79	cycle	89	temperature	99	absorbed
60	organs	70	majority	80	mature	90	essentially	100	category

Sub-list 3

101	clinic	111	functional	121	techniques	131	functioning	141	impact
102	efficient	112	method	122	network	132	inject	142	swollen
103	mechanical	113	capacity	123	tubes	133	lectures	143	appropriate
104	responses	114	task	124	extent	134	oral	144	aspect
105	equation	115	layer	125	joints	135	review	145	visible
106	personality	116	consultant	126	presence	136	critical	146	injection
107	potentially	117	elements	127	scan	137	defined	147	contrast
108	stimulate	118	interaction	128	aim	138	independent	148	decreased
109	academic	119	panel	129	assessed	139	infections	149	devices
110	candidates	120	proportion	130	density	140	essential	150	initially

Sub-list 4

151	moderate	161	reverse	171	criteria	181	tender	191	disorders
152	strategy	162	bacterial	172	excess	182	coordination	192	echo
153	content	163	elevated	173	focused	183	ratio	193	insert
154	emergency	164	interpret	174	hips	184	disc	194	objective
155	solve	165	monitor	175	insight	185	disorder	195	psychological
156	decrease	166	multiple	176	FALSE	186	expand	196	responding
157	error	167	swelling	177	consequences	187	outcomes	197	admission
158	minimum	168	aggressive	178	display	188	samples	198	assignment
159	predict	169	inserted	179	monitoring	189	tenderness	199	extensive
160	procedure	170	accurate	180	professor	190	conclusions	200	index

Sub-list 5

201	outcome	211	scans
202	structural	212	symptom
203	absence	213	intervention
204	clinically	214	orientation
205	collapse	215	persistent
206	confirm	216	sustained
207	disability	217	administrative
208	evaluate	218	fragment
209	underlying	219	sensitivity
210	experimental		

Level 4: MSWL words at the 1st 1,000 BNC/COCA word levels

Sub-list 1

1	kidney	11	plasma	21	potassium	31	cardiac	41	creatinine
2	renal	12	acute	22	GP	32	diabetic	42	contraction
3	urine	13	diabetes	23	lungs	33	syndrome	43	ET
4	liver	14	aorta	24	kidneys	34	fracture	44	surgeons
5	fluid	15	beta	25	pancreas	35	trauma	45	antibody
6	membrane	16	chronic	26	diagnosis	36	ultrasound	46	impairment
7	sodium	17	artery	27	glucose	37	marrow	47	cardiovascular
8	calcium	18	feedback	28	hypertension	38	vitamin	48	coronary
9	insulin	19	arthritis	29	arteries	39	GPS	49	alpha
10	oxygen	20	lung	30	antibodies	40	loop	50	proximal

Sub-list 2

51	acid	61	surgeon	71	necrosis	81	capillary	91	differential
52	duct	62	abnormal	72	physics	82	fluids	92	milligrams
53	fraction	63	activated	73	veins	83	inflammatory	93	fever
54	pulmonary	64	biopsy	74	respiratory	84	abdomen	94	fibrosis
55	enzyme	65	urinary	75	antibiotics	85	cartilage	95	cholesterol
56	femur	66	vascular	76	anatomy	86	femoral	96	prognosis
57	pelvis	67	cellular	77	metabolism	87	surgical	97	deficiency
58	albumin	68	membranes	78	abdominal	88	distal	98	fractures
59	enzymes	69	vein	79	biology	89	lymphoma	99	micro
60	particles	70	CT	80	metabolic	90	transplant	100	physicians

Sub-list 3

101	HIV	111	terminal	121	viral	131	pelvic	141	autoimmune
102	vasculitis	112	capsule	122	dye	132	spleen	142	cavity
103	hormone	113	gamma	123	lap	133	tumour	143	graft
104	pathology	114	grams	124	steroids	134	cortex	144	lesion
105	basement	115	peripheral	125	toxic	135	diagnose	145	obstruction
106	cord	116	tract	126	diagnostic	136	diagnosed	146	platelets
107	cirrhosis	117	vitro	127	intravenous	137	intake	147	sickle
108	physiological	118	diffuse	128	scar	138	minimal	148	BP
109	spectrum	119	pancreatic	129	hypertensive	139	pulse	149	clotting
110	systemic	120	spinal	130	lifetime	140	spine	150	lesions

Sub-list 4

151	microscope	161	transplants	171	onset	181	induced
152	physician	162	ward	172	ruptured	182	nausea
153	defect	163	bowel	173	anterior	183	sensory
154	enlarged	164	inhibitors	174	bilateral	184	deposition
155	registrar	165	jaundice	175	thigh	185	gram
156	cognitive	166	myocardial	176	infarction	186	toxins
157	thyroid	167	rotation	177	compression		
158	defects	168	serum	178	arterial		
159	inflammation	169	inhibitor	179	nodes		
160	malignant	170	lateral	180	sepsis		

Appendix C. Cumulative coverage including proper nouns and marginal words of the medical television program corpus and each program in the corpus

Table 1. Coverage for the whole medical television program corpus (%)

Word list	Coverage at each 1,000-word level	Cumulative coverage without proper nouns & marginal words	Cumulative coverage with proper nouns & marginal words
1,000	86.01	86.01	89.77
2,000	3.78	89.79	93.55
3,000	1.61	91.40	95.16*
4,000	1.07	92.47	96.23
5,000	0.67	93.14	96.9
6,000	0.44	93.58	97.34
7,000	0.3	93.88	97.64
8,000	0.26	94.14	97.90
9,000	0.19	94.33	98.09**
10,000	0.14	94.47	98.23
11,000	0.13	94.60	98.36
12,000	0.11	94.71	98.47
13,000	0.09	94.8	98.56
14,000	0.06	94.86	98.62
15,000	0.05	94.91	98.67
16,000	0.05	94.96	98.72
17,000	0.03	94.99	98.75
18,000	0.04	95.03*	98.79
19,000	0.02	95.05	98.81
20,000	0.03	95.08	98.84
21,000	0.01	95.09	98.85
22,000	0.01	95.1	98.86
23,000	0.01	95.11	98.87
24,000	0.01	95.12	98.88
25,000	0.01	95.13	98.89
Proper nouns	2.32		
Marginal words	1.44		
Off list	1.11		
Tokens	11,036,771		

*reaching 95% coverage, **reaching 98% coverage

Table 2- Coverage for 18 programs that need a vocabulary size of 3,000 words to reach 95% coverage

Word list	#5	#6	#7	#8	#9	#10	#11	#12	#13
1000	88.49	90.08	89.79	89.59	90.98	89.92	90.86	90.99	91.6
2000	93.22	93.8	94.22	93.4	94.68	93.54	94.36	94.38	94.96
3000	95.42*	95.41*	95.83*	95.02*	95.97*	95.41*	95.41*	95.92*	96.38*
4000	96.49	96.49	96.87	96.19	97.13	96.37	96.18	96.87	97.13
5000	97.16	97.08	97.52	96.8	97.66	97	96.83	97.53	97.74
6000	97.68	97.52	97.89	97.17	97.97	97.49	97.28	97.91	98.11**
7000	98.14**	97.81	98.18**	97.44	98.21**	97.71	97.62	98.12**	98.29
8000	98.38	98.05**	98.41	97.72	98.4	97.99	97.88	98.33	98.49
9000	98.5	98.21	98.58	97.96	98.55	98.1**	98.1**	98.5	98.63
10000	98.62	98.32	98.7	98.09**	98.66	98.23	98.27	98.63	98.71
11000	98.74	98.43	98.83	98.22	98.77	98.32	98.47	98.74	98.78
12000	98.88	98.56	98.93	98.33	98.84	98.42	98.61	98.83	98.85
13000	98.93	98.65	99	98.44	98.9	98.51	98.72	98.89	98.89
14000	98.96	98.7	99.04	98.5	98.94	98.6	98.8	98.94	98.92
15000	99	98.74	99.08	98.56	98.98	98.63	98.84	98.99	98.95
16000	99.01	98.8	99.12	98.61	99.01	98.71	98.88	99.04	98.98
17000	99.02	98.83	99.15	98.64	99.03	98.74	98.92	99.06	99
18000	99.08	98.86	99.18	98.66	99.07	98.76	98.96	99.08	99.02
19000	99.09	98.87	99.2	98.67	99.09	98.79	98.98	99.1	99.03
20000	99.16	98.95	99.22	98.7	99.1	98.83	99.00	99.13	99.05
21000	99.18	98.97	99.24	98.72	99.11	98.85	99.01	99.14	99.06
22000	99.19	98.98	99.25	98.74	99.12	98.86	99.03	99.16	99.07
23000	99.2	99	99.26	98.75	99.13	98.87	99.04	99.17	99.07
24000	99.2	99.01	99.26	98.76	99.13	98.88	99.05	99.18	99.09
25000	99.21	99.01	99.26	98.77	99.14	98.88	99.05	99.18	99.1
Size	14,666	95,542	233,702	257,221	152,463	49,917	109,156	1,885,037	95,165

Table 3- Coverage for Eighteen programs that need a vocabulary size of 3,000 words to reach 95% coverage (cont.)

Word list	#14	#15	#16	#17	#18	#19	#20	#21	#22
1000	91.2	90.38	91.15	89.68	90.4	90.45	88.54	89.74	91.13
2000	94.45	94.2	94.57	93.49	93.99	93.93	93.16	94.03	94.44
3000	95.68*	95.75*	95.78*	95.09*	95.51*	95.16*	95.1*	95.48*	95.71*
4000	96.58	96.72	96.7	96.07	96.61	96.01	96.23	96.49	96.72
5000	97.15	97.31	97.3	96.66	97.2	96.71	96.97	97.16	97.37
6000	97.6	97.71	97.73	97.08	97.58	97.24	97.35	97.49	97.73
7000	97.86	98**	97.97	97.32	97.88	97.5	97.71	97.72	97.95
8000	98.14**	98.26	98.22**	97.54	98.17**	97.79	97.95	97.96	98.2**
9000	98.34	98.41	98.49	97.72	98.36	97.97	98.23**	98.07**	98.42
10000	98.46	98.54	98.58	97.83	98.53	98.11**	98.49	98.23	98.53
11000	98.64	98.66	98.69	97.93	98.65	98.26	98.66	98.36	98.65
12000	98.74	98.74	98.77	98.05**	98.76	98.36	98.8	98.45	98.74
13000	98.81	98.92	98.84	98.12	98.83	98.42	98.88	98.55	98.83
14000	98.86	98.96	98.89	98.19	98.89	98.47	98.94	98.62	98.89
15000	98.9	99.00	98.92	98.23	98.93	98.52	98.99	98.68	98.92
16000	98.94	99.02	98.96	98.27	98.98	98.56	99.02	98.75	98.95
17000	98.96	99.05	98.98	98.30	99.01	98.59	99.07	98.78	98.98
18000	98.99	99.07	99.00	98.33	99.04	98.62	99.17	98.80	99.00
19000	99	99.1	99.02	98.35	99.07	98.67	99.18	98.82	99.02
20000	99.02	99.12	99.04	98.37	99.1	98.69	99.20	98.86	99.04
21000	99.03	99.13	99.05	98.39	99.12	98.7	99.21	98.89	99.05
22000	99.05	99.14	99.05	98.4	99.13	98.71	99.22	98.91	99.06
23000	99.06	99.14	99.05	98.41	99.16	98.72	99.24	98.93	99.07
24000	99.07	99.14	99.06	98.41	99.17	98.72	99.24	98.94	99.08
25000	99.07	99.14	99.06	98.43	99.18	98.73	99.24	98.95	99.09
Size	117,393	479,689	236,062	523,576	476,695	556,999	96,524	30,601	233,670

Table 4- Coverage for 4 programs that need a vocabulary size of 2,000 words to reach 95% coverage

Word list	#1	#2	#3	#4
1000	91.88	92.41	91.86	91.14
2000	95.39*	95.36*	95.31*	95.29*
3000	96.45	96.59	96.68	96.52
4000	97.38	97.26	97.6	97.56
5000	97.88	97.68	97.85	98.08**
6000	98.17**	97.95	98.26**	98.33
7000	98.45	98.27**	98.57	98.58
8000	98.67	98.43	99.11	98.77
9000	98.86	98.57	99.16	98.91
10000	98.95	98.65	99.24	99.02
11000	99.06	98.73	99.39	99.12
12000	99.13	98.84	99.43	99.18
13000	99.19	98.87	99.46	99.21
14000	99.23	98.9	99.46	99.25
15000	99.29	98.92	99.46	99.28
16000	99.31	98.94	99.46	99.31
17000	99.33	98.96	99.46	99.33
18000	99.35	98.97	99.46	99.34
19000	99.36	98.98	99.46	99.36
20000	99.38	98.99	99.46	99.36
21000	99.39	99	99.46	99.36
22000	99.4	99.01	99.46	99.36
23000	99.4	99.02	99.46	99.37
24000	99.4	99.02	99.46	99.37
25000	99.4	99.02	99.46	99.37
Size	161,600	648,399	39,506	92,121

Table 5. Coverage for 15 programs that need a vocabulary size of 4,000 words or more to reach 95% coverage

word list	#23	#24	#25	#26	#27	#28	#29
1000	89.86	88.36	88.38	89.07	88.48	87.19	88.24
2000	93.48	92.49	93.06	93.09	92.38	91.73	93.09
3000	94.8	94.48	94.76	94.91	94.12	93.99	94.78
4000	95.69*	95.78*	95.82*	96.3**	95.41**	95.33**	95.96**
5000	96.48	96.53	96.34	96.88	96.15	96.16	96.57
6000	96.9	96.96	96.86	97.32	96.65	96.74	97.12
7000	97.17	97.28	97.19	97.54	96.97	97.25	97.42
8000	97.5	97.56	97.43	97.76	97.26	97.55	97.64
9000	98**	97.77	97.74	98.01**	97.45	97.78	97.82
10000	98.27	97.91	97.96	98.13	97.6	97.98	98.01**
11000	98.35	98.05**	98.08**	98.36	97.76	98.12**	98.19
12000	98.47	98.18	98.2	98.47	97.89	98.25	98.36
13000	98.70	98.28	98.28	98.57	98.03**	98.36	98.47
14000	98.77	98.37	98.39	98.63	98.12	98.44	98.51
15000	98.80	98.44	98.5	98.66	98.19	98.51	98.55
16000	98.82	98.48	98.55	98.69	98.25	98.56	98.62
17000	98.91	98.52	98.62	98.74	98.3	98.6	98.66
18000	98.91	98.55	98.64	98.79	98.33	98.73	98.68
19000	99.04	98.58	98.65	98.81	98.36	98.76	98.7
20000	99.05	98.61	98.67	98.86	98.41	98.79	98.73
21000	99.07	98.62	98.67	98.89	98.43	98.81	98.74
22000	99.07	98.63	98.67	98.9	98.46	98.83	98.75
23000	99.07	98.64	98.67	98.9	98.48	98.85	98.76
24000	99.07	98.65	98.67	98.91	98.51	98.86	98.77
25000	99.07	98.65	98.68	98.92	98.52	98.86	98.8
Size	9,015	443,563	33,153	57,221	1,828,602	932,701	42,490

Table 6. Coverage for 15 programs that need a vocabulary size of 4,000 words or more to reach 95% coverage (cont.)

Word list	#30	#31	#32	#33	#34	#35	#36	#37
1000	88.75	88.53	88.53	89.23	89.06	87.67	86.98	86.23
2000	93.24	92.62	92.62	93.22	92.51	91.95	91.54	90.65
3000	94.47	94.72	94.72	94.65	93.86	94.29	94.15	92.54
4000	95.85*	95.76*	95.76*	95.72*	95.05*	95.49*	95.42*	94.39
5000	96.92	96.43	96.43	96.47	95.85	96.26	96.16	95.28*
6000	97.59	96.9	96.9	97	96.26	96.72	96.74	95.73
7000	97.79	97.21	97.21	97.34	96.57	97.03	97.09	96.06
8000	98.11**	97.34	97.34	97.64	96.81	97.38	97.38	96.39
9000	98.31	97.67	97.67	97.87	97.07	97.57	97.55	96.7
10000	98.38	97.85	97.85	98.06**	97.2	97.73	97.71	96.93
11000	98.53	97.96	97.96	98.21	97.34	97.89	97.87	97.09
12000	98.62	98**	98**	98.34	97.46	98.06**	97.99	97.22
13000	98.71	98.15	98.15	98.44	97.54	98.13	98.08**	97.59
14000	98.74	98.27	98.27	98.52	97.6	98.22	98.15	97.7
15000	98.82	98.3	98.3	98.59	97.65	98.28	98.18	97.75
16000	98.85	98.36	98.36	98.65	97.73	98.37	98.29	97.8
17000	98.88	98.39	98.39	98.7	97.79	98.41	98.31	97.81
18000	98.91	98.4	98.4	98.73	97.88	98.46	98.34	97.84
19000	98.93	98.45	98.45	98.76	97.89	98.48	98.39	97.85
20000	99.04	98.46	98.46	98.78	97.93	98.53	98.42	97.88
21000	99.05	98.46	98.46	98.79	97.95	98.56	98.44	97.9
22000	99.05	98.47	98.47	98.8	97.96	98.57	98.46	97.93
23000	99.05	98.48	98.48	98.81	97.97	98.58	98.47	97.93
24000	99.05	98.48	98.48	98.81	97.97	98.59	98.49	97.97
25000	99.05	98.49	98.49	98.82	97.98	98.6	98.5	97.98
Size	81,058	9,428	61,565	623,500	67,940	116,290	112,593	31,948