A Corpus Comparison Approach for Estimating the Vocabulary Load of Medical Textbooks Using The GSL, AWL, and EAP Science Lists

Betsy Quero*
Victoria University of Wellington, New Zealand

Abstract

The main goal of this study is to report on the number of words (vocabulary load) native and non-native readers of medical textbooks written in English need to know in order to be able to meet the lexical demands of this type of subject-specific (medical) texts. For estimating the vocabulary load of medical textbooks, a corpus comparison approach and some existing word lists, popular in ESP and EAP, were used. The present investigation aims to answer the following questions: (1) How many words are needed beyond the General Service List (GSL; West, 1953), the Academic Word List (AWL; Coxhead, 2000), and the EAP Science List (Coxhead and Hirsh, 2007) to achieve a good lexical text coverage? and (2) What is the vocabulary load of medical textbooks written in English? The implementation of this corpus comparison approach consisted of: (1) making a written medical corpus of 5.4 million tokens, (2) compiling a general written corpus of the same size (5.4 million tokens), (3) running both corpora (i.e., the medical and general) through some existing word lists (i.e., the GSL, the AWL, and the EAP Science List), and (4) creating new subject-specific (medical) word lists beyond the existing word lists used. The system for identifying medical words was based on Chung and Nation's (2003) criteria for classifying specialised vocabulary. The results of this investigation showed that there is a large number of subject-specific (medical) words in medical textbooks. For both native and non-native speakers of English training to be health professionals, this figure represents an enormous amount of vocabulary learning. This paper concludes by considering the value of creating specialised medical word lists for research, teaching and testing purposes.

Key words: medical word lists, vocabulary load, English for medical purposes, text coverage.

Introduction

One of the main purposes of this study is to propose a methodology for the creation of subject-specific word lists (i.e., medical word lists) that include the most salient vocabulary in medical texts. After doing a review of the previous studies on the vocabulary load of medical textbooks, explaining the methodology and presenting the subject specific lists of the most relevant words in medical texts, the results of this investigation attempt to: (1) identify the lexical demands of medical texts using a corpus comparison approach, and (2) provide guidelines for the creation of medical word lists organised by levels of frequency and salience.

Vocabulary Load

The number of known words (vocabulary load) needed for unassisted reading comprehension has been investigated by several vocabulary researchers (Hirsh & Nation, 1992; Hu & Nation, 2000; Laufer, 1989; Nation, 2006). The first investigations (Laufer, 1989, 1992) on the vocabulary load of academic texts suggested a reading

* *Tel: + 64 2102387831; E-mail: betsy.quero@vuw.ac.nz; PO Box 14416 Kilbirnie, Wellington 6241, New Zealand*

comprehension threshold of 95% text coverage. More recent research on the vocabulary load of written texts (Hu and Nation 2000; Laufer and Ravenhorst-Kalovski 2010; Nation 2006; Schmitt, Jiang, and Grabe 2011) has indicated that a higher lexical threshold of 98% text coverage or more is required for optimal unassisted reading comprehension. In the present study, we explore the number of words required to be known to achieve a 98% text coverage, and refer to 98% as an optimal lexical threshold.

Levels of Vocabulary

In order to estimate the number of words (vocabulary load) that learners of English for Medical Purposes (EMP) need to know in order to be able to meet the vocabulary demands of medical texts written in English and achieve a suitable reading comprehension threshold (i.e., between 95% and 98% text coverage); the various levels of vocabulary proposed by Schmitt and Schmitt (2012) and Nation (2001, 2013) will be identified in the corpus of medical textbooks compiled for this study. Frequency (high-frequency, mid-frequency, and low-frequency words), and text type (i.e., general, academic, scientific, technical or specialised) are the two main criteria currently used to classify the vocabulary of academic and specialised texts.

Schmitt and Schmitt's (2012) classification of the levels of vocabulary is a frequency-based one, and consists of the following three bands or levels: high-frequency, mid-frequency, and low-frequency words. The high-frequency level includes the first 3,000 most frequent words in a language. The mid-frequency level refers to those words between the 4,000 and the 9,000 frequency levels. The low-frequency level comprises those words beyond the 9,000 frequency band. The concept of mid-frequency vocabulary was first introduced in Schmitt and Schmitt's (2012) classification. The introduction of this frequency level has served to stress the importance of mid-frequency vocabulary and of words beyond the 3,000 most frequent words of the English language.

Nation's (2013) classification, which was initially presented in 2001 and then revised in 2013, is both a frequency and text-type based classification. Nation's (2001) frequency levels included two frequency bands (i.e., high-frequency vocabulary and low-frequency vocabulary) and two kinds of text type words (academic vocabulary and technical vocabulary). In 2013 Nation added to his classification of vocabulary levels the mid-frequency band proposed by Schmitt and Schmitt in 2012. According to Nation (2013), there are three levels of frequency based words, that is, high-frequency words, mid-frequency words and low-frequency words, and two levels of text-type words (academic words and technical words) which are particularly likely to occur in academic and specialised texts. Both the frequency and text-type based aspects of Nation's (2013) classification are analysed and discussed in the findings and discussion sections of this study.

Word Lists in EAP and ESP

High-frequency general, academic and specialised word lists have been used in English for Academic Purposes (EAP) and English for Specific Purposes (ESP) by language teachers, students, researchers, test designers, and course material developers. To the best of our knowledge, the most extensively used and discussed high-frequency general academic word lists in EAP and ESP have been West's (1953) General Service List (GSL) and Coxhead's (2000) Academic Word List (AWL). More recently, Coxhead and Hirsh (2007) developed an EAP Science List that was created excluding words in the GSL and the AWL.

West's (1953) General Service List (GSL) is a high-frequency list of English words that contains roughly 2,000 words (i.e., GSL1 with the first 1,000 and GSL2 with the second 1,000 most frequent word families) which are very common in all uses of the language. For more than 60 years, the GSL has been the most widely used high-frequency word list for language curriculum planning, materials development, and vocabulary instruction. The GSL has been criticised for its age (Hyland & Tse, 2007; Read, 2000, 2007), for its size (Engels, 1968), and for its lack of suitability to the vocabulary needs of ESP learners at tertiary level (Ward, 1999, 2009). For decades, vocabulary researchers constantly stated that the GSL was in need of revision (Coxhead, 2000; Hwang & Nation, 1989; Wang & Nation, 2004); however, it was not until its 60th anniversary that two new general vocabulary lists (Brezina & Gablasova, 2013; Browne, 2013) were created. Despite the criticism West's (1953)

GSL has received over the years, this is the general word list used in this study to replicate the corpus comparison approach. The GSL is used in this investigation in order to: (1) serve as a starting point when estimating the vocabulary load of medical texts, and (2) allow comparisons with previous studies in ESP that have also used the GSL to look at the number of words in the health and medical sciences.

The other existing word list used in the present study is Coxhead's (2000) Academic Word List (AWL). The AWL works in conjunction with the GSL. That is, it includes words that do not occur in the GSL. Up to the present, the AWL has been extensively used to learn, teach, and research academic vocabulary. To make the AWL, Coxhead (2000) gathered a corpus of 3,513,330 tokens. This corpus was comprised of a variety of academic texts from 28 academic subject areas, seven of which were grouped into one of the following four disciplines: Arts, Commerce, Law, and Science. The AWL contains 570 word families and provides around a 10% text coverage for academic texts. For validating the AWL, Coxhead (2000) created a second academic corpus (comprising 678,000 tokens) which accounted for 8.5% coverage.

Two new academic word lists have been recently developed: (1) The New Academic Word List (NAWL) created by Browne, Culligan, and Phillips in 2013 and available at http://www.newacademicwordlist.org/, and (2) The New Academic Vocabulary List (AVL) created by Gardner and Davies (2014) and available at http://www.academicvocabulary.info/download.asp. Both the NAWL and the AVL were developed from large academic corpora of 288 and 120 million tokens, respectively. Despite the current availability of these more recently developed academic word lists (i.e., the NAWL and the AVL), the decision to use Coxhead's (2000) AWL for the present study is based on the fact that for more than a decade the AWL has been widely researched and used by ESP researchers to calculate the lexical demands posed by written academic texts.

Drawing on some aspects of the methodology used by Coxhead (2000) to create the AWL, various subject-specific word lists have been developed: an EAP Science Word List (Coxhead & Hirsh, 2007), three medical academic word lists (Chen & Ge, 2007; Lei & Liu, 2016; Wang, Liang, & Ge, 2008), a nursing word list (Yang, 2015) a pharmacology word list (Fraser, 2007), some engineering word lists (Mudraya, 2006; Ward, 1999, 2009), a business word list (Konstantakis, 2007), and an agricultural word list (Martínez, Beck, & Panza, 2009). While some of these subject-specific lists have been developed to work in conjunction the GSL (e.g., Yang's (2015) Nursing Word List, and Wang, Liang & Ge's (2008) Medical Academic Word List), other word lists have been created to work in conjunction with both the GSL and AWL (e.g., Coxhead and Hirsh's (2007) EAP Science List, and Fraser's (2007) Pharmacology Word List).

Coxhead and Hirsh's (2007) EAP Science List is another existing word list used in the present study to estimate the vocabulary load of medical textbooks. Coxhead and Hirsh's (2007) study aims at creating a science word list that could help increase the lower coverage of the AWL over science texts (Coxhead, 2000). Criteria of range, frequency of occurrence, and dispersion were considered for selecting the words to be added to the EAP Science List. This list is based on a written science corpus of English comprising a total of 2,637,226 tokens. As Coxhead and Hirsh (2007, p. 72) reported, the 318 word families in the EAP Science List cover 3.79% over the science corpus compiled to create this list. Moreover, the EAP Science list covers 0.61% over the Arts subcorpus, 0.54% over the Commerce subcorpus, 0.34% over the Law subcorpus, and 0.27% over the fiction corpus compiled by Coxhead (2000). The above mentioned coverage results confirm the scientific nature of the EAP Science List. Coxhead and Hirsh's (2007) study also attempts to draw a line between the percentage of general vocabulary versus the percentage of science-specific vocabulary in science texts written in English that EAP students are required to read at university. In addition to the GSL and the AWL, Coxhead and Hirsh's (2007) EAP Science List is used in the present investigation when adopting the corpus comparison approach to estimate the vocabulary load of medical textbooks.

Since the present study focuses on investigating the vocabulary load of the most commonly used existing general, academic and scientific word lists, these lists are used as the starting point to estimate the lexical coverage of medical texts. By choosing a set of commonly used general/academic/scientific word lists, this study tries to focus on general/academic/scientific vocabulary that has extensively been presented in EAP and ESP teaching materials, assessments, and research. However, this investigation by no means attempts to undermine the value of more recently created general (i.e., the two NGSLs) and academic (i.e., the NAWL and the AVL)

word lists. Also, to the best of our knowledge, no study has so far estimated the vocabulary load of medical textbooks having as a starting point for this quantification this set of widely used word lists (i.e., the GSL, the AWL, and the EAP Science List) in EAP and ESP.

Moreover, existing pedagogical vocabulary lists of general high-frequency words (West's GSL) and academic words (Coxhead's AWL), and scientific words (Coxhead and Hirsh's EAP Science List) cannot provide a complete coverage of the kinds of vocabulary in subject-specific texts. This happens particularly because the GSL, the AWL and the EAP Science List were not designed to identify all the different kinds of vocabulary of specialised texts. For this reason, a more inclusive approach to identify the various levels of vocabulary that occur in medical texts could provide a clearer picture of the vocabulary demands of medical textbooks.

Research Questions

The present investigation looks at the vocabulary load of medical texts and explores the role played by the levels of vocabulary proposed by Nation (2013) and Schmitt and Schmitt (2012). In particular, the three frequency-based levels of vocabulary (high, mid, and low-frequency words) and four topic-based word lists (the GSL, the AWL, the EAP Science List, and some specialised medical lists) that draw on words from these three frequency levels were used in the analyses of the lexical frequency profiles of medical texts here investigated. With the main goal of estimating the vocabulary load of medical textbooks in mind, the findings of this study provide answers to the following research questions:

1) How many words are needed beyond the General Service List (GSL; West, 1953), the Academic Word List (AWL; Coxhead, 2000), and the EAP Science List (Coxhead and Hirsh, 2007) to achieve a good lexical text coverage?

2) What is the vocabulary load of medical textbooks written in English?

Methodology

The methodology used to estimate the number of words (vocabulary load) associated with the various levels of vocabulary found in a corpus of medical textbooks is discussed in this section. The implementation of this methodology involves compiling the medical and general corpora, adopting a corpus comparison approach, adapting a semantic rating scale, creating a series of medical word lists, and justifying the unit of counting selected for the present study.

Compiling the Corpora

The estimation of the vocabulary load of medical textbooks using a corpus comparison approach required the use of two different corpora: a specialised (medical) corpus and a general corpus. For the medical corpus, two widely consulted handbooks of general medicine were selected (i.e., Harrison's Principles of Internal Medicine by Fauci et al., 2008, and Cecil Textbook of Internal Medicine by Goldman & Ausiello, 2008). These two medical textbooks include a comprehensive range of medical topics, and are commonly consulted by both medical students (from the first year of medical studies) and health professionals. In relation to the general corpus created to serve as a general comparison corpus for this study, it was compiled using most sections of seven general English corpora, namely the FLOB corpus (British English 1999), FROWN corpus (American English 1992), KOLHAPUR corpus (Indian English 1978), LOB corpus (British English 1961), WWC corpus (New Zealand English 1993), BROWN corpus (American English 1961), and ACE corpus (Australian English 1986). Only section J (i.e., the learned section) was removed from all the general corpora used before compiling them. Both the medical and general corpora are the same size (5,431,740 tokens each) so that distortion from adjusting for various corpus sizes could be avoided when using the corpus comparison approach.

Adopting a Corpus Comparison Approach

The use of the corpus-comparison approach involved largely following Chung and Nation's (2003) procedure to find potential technical vocabulary. Corpus-comparison entails the use of two different corpora: a non-technical corpus and a technical corpus to compare word frequencies. Moreover, the corpus comparison procedure involves comparing word frequencies in two corpora and choosing words that are much more frequent in the technical corpus than in a non-medical comparison corpus, or that are unique to the technical corpus, as potential technical words. Also, words occurring only in the technical corpus or with a higher frequency in the technical corpus are more likely to be technical words. Range (Heatley, Nation, & Coxhead, 2002) was the software used to carry out the frequency comparison of the medical and the general corpora.

Adapting a Semantic Rating Scale

As part of the procedures followed in this study to classify medical words and estimate the lexical demands of medical textbooks, Chung and Nation's (2003, 2004) methodology for identifying content area (technical) words was used. Their methodology is twofold, involving the use of a semantic rating scale, and a corpus comparison approach. Here we propose a semantic rating scale for classifying words related to health and medicine. This rating scale approach for identifying medical words will be combined with a corpus-based approach for looking at technical words in medical texts. These potential technical words needed to be checked systematically to decide if they were truly technical words. This required the development of a checking system.

Only a yes/no decision-making procedure was required to decide whether words were to be considered as medical or not, therefore the classification needed to use a semantic rating scale with two levels, namely, general purpose vocabulary versus content area (technical) vocabulary. To guide this decision-making, four sub-levels of medical words were used to ensure consistency. The starting point for the system was Chung and Nation's (2003) rating scale which was originally designed to identify the specialised vocabulary used in anatomy and applied linguistic texts. The adaptation of Chung and Nation's (2003) rating scale for this study consisted of grouping the four levels of their semantic rating scale into two main levels to classify vocabulary into (1) general purpose vocabulary and (2) content area (technical) vocabulary. Meaning is the main feature used to classify the vocabulary in medical texts according to the rating scale developed for the present study. The primary purpose of the semantic rating scale is to draw the line between (1) general purpose vocabulary, and (2) content area (technical) vocabulary in medical texts written in English. The four sub-levels of content area (medical) vocabulary for the present study are as follows:

- Sub-level 1: Some topic-related words are also general purpose words used in the medical field with the same meaning they most frequently have in other general fields and everyday usage. Examples are words such as *nurse, doctor, child, medicine, blood, pain, health*.
- Sub-level 2: Some topic-related words are general purpose vocabulary used in the medical field, but with a particular meaning not so frequently encountered in general fields and everyday usage. Examples are words such as *transcription, pressure, antagonists*.
- Sub-level 3: Some topic-related words are associated with more than one particular specialised subject area with the same meaning. An expert in this particular field where these words come from would identify these words as words specific to their discipline. Examples are words such as *nitrogen, ethanol, fluorine* from Chemistry; and *species, organisms, nature* from Biology. These words are also used to talk and write about health and medicine.
- Sub-level 4: Some topic-related words are unique to the medical field, and they are only associated with highly specialised medical topics. These medical words have a subject-specific meaning, and are very unlikely to be found in other disciplines. That is, they will only or almost exclusively be used within the medical field. An expert in the medical and health sciences can identify them as technical or scientific words specific to the subject area. Examples of highly technical words in the medical field are *schistosomiasis, polycythemia, dermatomyositis, enteropathy* and *hemochromatosis*. These highly specialised medical words are most likely to be only known by specialists in the medical and health sciences.

These criteria were also used to decide whether the words from the GSL, AWL, EAP Science List and medical word lists were general or medical words. Manual checking was used to identify the words found by

corpus comparison. General words referring to abbreviations, living organisms, parts of the body, participants in the health and medical community were classified as medical words. The manual checking of all the word types (including content words, abbreviations, acronyms and proper nouns) classified using the semantic rating scale involved: (1) looking up word types with unclear medical meaning in a specialised medical dictionary and (2) confirming the medical senses of these in their actual context of occurrence in the medical corpus.

Developing New Medical Word Lists Through Corpus-Comparison

The General Service List (West, 1953), the Academic Word List (Coxhead, 2000), and the EAP Science List (Coxhead & Hirsh, 2007) were used because the words in these lists are assumed to already be known by first and second year medical students. The words not found in any of the lists were organised and classified following two different procedures to make medical word lists: one with the words occurring in both corpora using frequency comparison, and one with the words occurring only in the medical corpus. That is, the creation of the medical word types was done by: (1) choosing the most frequent 3,000 medical word types occurring in both the medical and general corpora and ranking first the word types with higher relative frequency (medical frequency of each word type divided by general frequency of the same word type), and then (2) selecting 23,000 unique medical word types and ranking the word types by their absolute frequency of occurrence in the medical corpus. These 26 new medical word lists included only word types that have been previously classified as medical words using the yes/no decision-making procedure developed using the semantic rating scale previously mentioned. We decided it would be better to keep the medical words occurring in both the medical and general corpora, and the medical words unique to the medical corpus in separate word lists. The rationale behind this decision is twofold: (1) ranking the two kinds of word lists separately provides better coverage with a smaller amount of word types than ranking them together, and (2) these two kinds of medical words may involve different learning procedures.

Selecting the Unit of Counting

The decision about which unit of counting to use (word types, lemmas or families) depends on the goals of the study and beliefs about relationships between lemma and word family members (see Nation, 2016 for further discussion on units of counting). The word type is the unit of counting selected in the present study to create new medical word lists, discuss the findings and estimate the vocabulary load of medical textbooks. The decision to use the word type as the unit of counting for this study was made because even though the GSL, AWL and EAP Science List word family members share the same core meaning, some word family members belong to different word classes and only one word type has a technical meaning. This happens because these lists (i.e., the GSL, AWL and EAP Science List) were made without grouping the word family members into word classes and taking meaning into consideration. Examples of word types belonging to different word classes and having different meanings in general and medical English are words such as *culture*, *patient*, and *radical*.

Results

How many words are needed beyond the General Service List (GSL; West, 1953), the Academic Word List (AWL; Coxhead, 2000), and the EAP Science List (Coxhead and Hirsh, 2007) to achieve a good lexical text coverage?

This question is answered by presenting the cumulative text coverage results of running three sets of word lists: (Set 1) the GSL, AWL, EAP Science List, (Set 2) the three 1000 MGEN lists, and (Set 3) the twenty-three MED lists through the medical corpus using the Range software (Heatley et al., 2002). First, the cumulative coverage of the GSL1 and GSL2, the AWL and the EAP Science List, and the words outside these lists is presented in Table 1. Then, the cumulative text coverage of these three sets of word lists is summarised in Table 2.

Table 1 suggests that a 22.12% of the words outside the lists (i.e., the GSL1 and GSL2, AWL, and EAP Science List) is still needed to achieve an optimal lexical threshold of 98% (i.e., 75.88% coverage of word types in the lists plus 22.12% coverage of word types outside the lists). In order to find out how many more word types are required beyond the four existing word lists summarised in Table 1, we applied the semantic rating scale

described in the methodology section of the present study. This rating scale served as a semantic checking system to classify over 30,000 medical word types (see Quero, 2015) occurring in the medical corpus and create the 26 medical word lists whose text coverage results are summarised in Table 2.

Table 1
Cumulative Coverage of the GSL1 and GSL2, the AWL and the EAP Science List over the Medical Corpus including the Words outside the Lists

Word List	Coverage %	Number of Word Types
GSL1, GSL2, AWL, EAP Science List	75.88	9,412
Words outside the lists	24.12	45,942
Total	100.00	55,354

Table 2
Cumulative Coverage of the GSL, the AWL, the EAP Science List, the three 1,000 MEDGEN Lists, and the Twenty-three 1,000 MED Lists

Word List	Number of Tokens	Coverage %	Number of Word Types
GSL1, GSL2, AWL, EAP Science List	4,121,539	75.88	9,412
MGEN (three 1,000) lists	607,498	11.18	3,000
MED (twenty-three 1,000) lists	542,747	10.00	23,000
Cumulative total of existing lists	5,271,784	97.06	35,414

Note in Table 2 that the cumulative text coverage of the GSL1, GSL2, AWL and EAP Science List (75.88%) indicates that an additional 21.18% coverage is required to achieve a 97.06% text coverage. Moreover, the results in Table 2 show that 26,000 new medical word types (i.e., 3,000 medical word types in the MGEN lists, and 23,000 medical word types in the MED lists) need to be added to the GSL, AWL, and EAP Science List for readers of medical texts to be able to understand 97.06% of the words they meet when they read medical textbooks in English.

What is the vocabulary load of medical textbooks written in English?

The answer to this question is approached by looking at the behaviour of the three sets of word lists above mentioned, namely, the GSL, the AWL, and the EAP Science list (set 1), the three 1,000 MGEN word lists (set 2), and the twenty-three 1,000 MED lists (set 3).

We start by looking at the text coverage results of the existing lists (i.e., the GSL, AWL, and EAP Science List). Then, we present the text coverage of the twenty-six 1,000 new medical word lists (i.e., the MGEN and MED lists) created for this study. In this section, the results of the coverage and frequency of occurrence of the word types across the GSL, AWL, EAP Science List, MGEN lists, MED lists and the words outside these lists are summarised in Tables 3, 4, 5 and 6.

As shown in Table 3, the GSL accounts for 55.62% of the medical corpus. Because the GSL1 includes the most frequent words of the English language and comprises the highest text coverage of the tokens in medical texts, this is a list of words worth learning for students of medical English. Regarding the lexical coverage of the GSL2, this list accounts for 5.97% of the tokens in the medical corpus. In general, it may be worth highlighting to medical students which are the medical words in the GSL that occur most frequently in medical textbooks. For instance, using the semantic rating scale described in the methodology section of this study, we identified 626 medical word types (out of a total of 4,119 word types) in the GSL1, and 371 medical word types (out of a total of 3,708 word types) in the GSL2. Examples of medical word types in the GSL are *bleeding*, *stroke*, and *illness* in the GSL1 and *health*, *pain,* and *brain* in the GSL2.

Table 3
Coverage of the GSL1 and GSL2, the AWL and the EAP Science List over the Medical Corpus

Word List	Coverage %	Number of Word Types
GSL1	55.62	3,291
GSL2	5.97	2,415
AWL	8.23	2,418
EAP Science List	6.06	1,288
Cumulative total	75.88	9,414

In relation to the coverage of the AWL over medical texts, Table 3 shows that the AWL accounts for 8.23% of the 5.4 million tokens of the medical corpus. 527 of the 3,107 word types in the AWL were identified as medical words. Examples of medical words in the AWL include *depression*, *labour*, and *topical*. When compared with the coverage of the GSL1 over medical texts, the 8.23% coverage of the AWL seems a good coverage of academic words over medicine. Since the lexical coverage by the AWL is 2.26% higher than that of the GSL2, these coverage results suggest that it may be more useful for ESP medical students to start learning the AWL right after they have acquired the words in the GSL1. The AWL is a particularly useful word list to learn when ESP medical students need to focus on academic words. For this reason, the AWL is a helpful list for medical students taking first year ESP reading courses.

As also indicated in Table 3, the high coverage of the EAP Science List over medicine (6.06%), when compared with the coverage of the GSL1, GSL2, and the AWL over medical texts, shows that EAP Science List plays an important complementary role in helping ESP medical students become familiar with scientific words that occur in texts of health and medicine (see Coxhead & Quero, 2015, for further discussion on the behaviour of the EAP Science List over medical texts). Examples of some scientific words with a medical meaning in the EAP Science List are *cell, anatomy,* and *digest.* These results also suggest that the EAP Science List is of particular interest to science and medical students rather than to learners of general English. Additionally, the lexical coverage results of the GSL, AWL and EAP Science List over the medical corpus suggest that the learning of high frequency general, academic and scientific words in English could be sequenced differently for ESP medical students.

Table 4
Cumulative Coverage of the Three 1,000 MGEN Lists

Word List	Coverage %	Number of Word Types
MGEN1	8.49	1,000
MGEN2	1.82	1,000
MGEN3	0.87	1,000
Cumulative total	11.18	3,000

Let us now look at the text coverage of the new medical word lists (i.e., the three 1,000 MGEN lists). These 3,000 medical word types are divided into three 1,000 word lists and referred to as MGEN1, MGEN2, and MGEN3 in Table 4. Examples of medical words in the MGEN lists are *syndromes*, *radiologist*, and *anatomical*. Note also in Table 4 that the three 1,000 MGEN lists provide a coverage of 11.18%. This means that the GSL, AWL, EAP Science List and the three MGEN list together cover 87.06% (i.e., 75.88% for the GSL, AWL and EAP Science List, and 11.18% for the three MGEN lists) of medical texts. This cumulative coverage of 87.06% indicates that a 10.94% coverage is still needed to reach an optimal lexical threshold of 98%.

Table 5 gives the coverage details of the twenty-three frequency-ranked 1,000 MED word lists that are unique to the medical corpus. As can be observed in Table 5, there is a large amount of low-frequency medical words occurring in medical texts. Examples of medical words in the 23 MED lists are *subcutaneously*, *polyarteritis*, and *catarrhalis*.

Table 5
Coverage of the Twenty-three 1,000 MED Lists

Word List	Coverage %	Number of Word Types
MED1	5.16	1,000
MED2	1.46	1,000
MED3	0.82	1,000
MED4	0.54	1,000
MED5	0.39	1,000
MED6	0.30	1,000
MED7	0.23	1,000
MED8	0.18	1,000
MED9	0.15	1,000
MED10	0.12	1,000
MED11	0.10	1,000
MED12	0.09	1,000
MED13	0.07	1,000
MED14	0.06	1,000
MED15	0.06	1,000
MED16	0.05	1,000
MED17	0.04	1,000
MED18	0.04	1,000
MED19	0.04	1,000
MED20	0.04	1,000
MED21	0.02	1,000
MED22	0.02	1,000
MED23	0.02	1,000
Cumulative total	10.00	23,000

Table 6 shows that 2.94% of the tokens and 19,942 word types occur in the medical corpus but not in the 30 existing word lists. These words outside the lists include single letters of the alphabet or roman numerals, marginal medical words (e.g., *chap* an abbreviation of chapter), prefixes (e.g., *non-*, and *micro-*), and low-frequency medical words (e.g., *encephalographic*, and *haematologist*).

Table 6
Coverage of the GSL, the AWL, the EAP Science List, the Three 1,000 MEDGEN Lists, and the Twenty-three 1,000 MED Lists Including Words outside the Existing Lists

Word List	Number of Tokens	Coverage %	Number of Word Types
Cumulative total of existing lists	5,271,784	97.06	35,414
Words outside the lists	159,956	2.94	19,942
Total	5,431,740	100.00	55,354

The cumulative coverage of all the 30 existing lists (i.e., the GSL, AWL, EAP Science List, and the three MGEN lists, and the twenty-three MED lists) and words outside these lists is compared in Table 6. The results in Table 6 show that if readers of medical texts want to get closer to a 98% text coverage over medical texts, a large number of the 19,942 word types left outside all these 30 word lists are required to achieve a 98% coverage. Based on the cumulative total coverage (97.06%) of the word lists shown in Table 6, we conclude that at least

twenty-two 1,000 low-frequency medical word lists would need to be added to these already existing 30 lists to increase the text coverage from 97.06% to 97.50% and start getting closer to 98% (the optimal lexical threshold). Another way to get closer to 98% with a smaller amount of word types could be to add word lists of high and mid-frequency words with general academic meaning that, for different reasons, are not included in the existing general, academic, and scientific word lists (i.e., the GSL, AWL, EAP Science List) used as part of the present investigation. (See also Appendix A with text coverage and occurrence figures of all the lists discussed in this study).

Discussion

Next, we discuss the value of the twofold methodology here adopted for identifying medical words. This discussion refers to the following aspects of the present study: (1) the semantic rating scale, (2) the size of the corpus, (3) the corpus comparison approach, and (4) the new medical word lists.

The Semantic Rating Scale

The replication of Chung and Nation's (2003) semantic rating scale involved the identification of thousands of medical words (over 30,000) occurring in the medical corpus used. Despite the usefulness of the semantic rating scale for making decisions on the number of content area vocabulary items found in medical texts, the need to classify thousands of words made the use of this rating scale a very time-consuming process (as also reported by Chung & Nation, 2004; Fraser, 2005, 2006). Likewise, there were still over 8,000 word types, most of them words occurring only once, that remained unclassified. The adaptation of Chung and Nation's (2003) rating scale for the present study has enabled u s to provide a comprehensive account of the lexical demands of medical textbooks. Hence, the use of Chung and Nation's semantic rating scale has proven effective to identify a large amount of words with medical meaning in the medical corpus – occurring in existing word lists such as West's (1953) GSL, Coxhead's (2000) AWL, Coxhead and Hirsh's (2007) EAP Science List, and the 26 medical word lists (i.e., the three 1,000 MGEN, and the twenty-three MED lists). In sum, the use of Chung and Nation's rating scale made possible the identification of a large number of content area (medical) words found in medical textbooks.

The Size of the Corpus

The size of the medical corpus was determined by the amount of specialised texts from a variety of medical topics available in digital format. The presence of a wide range of medical topics in the medical corpus facilitated the estimation of the lexical demands of medical textbooks. In fact, the size of the medical corpus was large enough in number of running words (5,431,740 tokens) and coverage of medical topics to estimate the lexical profile of medical texts and provide a representative sample of the lexis found in medical textbooks.

The Corpus Comparison Approach

As previously mentioned in the methodology section, two corpora (i.e., a medical corpus and a general corpus) were compiled to enable the implementation of the corpus comparison approach. These two corpora were characterised by having the same size (i.e., 5,431,740 tokens), but comprising different topics, namely, a variety of health and medical topics in the medical corpus, and a wide range of general topics in the general corpus. First of all, the medical corpus was created for identifying medical vocabulary, using Chung and Nation's semantic rating scale, in popular existing word lists – such as West's (1953) GSL, Coxhead's (2000) AWL, and Coxhead and Hirsh's (2007) EAP Science List – and beyond these lists. Then, the general comparison corpus of the same size was compiled to apply the corpus comparison approach for creating the medical word lists needed to estimate the lexical demands of medical textbooks. The use of two corpora (i.e., the medical and general corpora) of the same size but very different in their range of topics made possible the successful implementation of the corpus comparison approach for estimating the 98% lexical threshold of medical textbooks in this study.

The New Medical Word Lists

A series of medical word lists were created using two different frequency-based procedures. These two procedures were used to rank and group the medical words previously classified using an adaptation of Chung and Nation's (2003) semantic rating scale. The first procedure included medical words occurring both in the medical and general corpora: a total of three 1,000 MGEN word lists beyond the GSL, AWL, and EAP Science List were created applying this first procedure. The sets of medical word lists created following this first procedure were ranked by placing the medical word types with the highest relative frequency – which was calculated by dividing the frequency of a word type in the medical corpus by the frequency of the same word in the general corpus – at the top of the lists. The relative frequency, instead of the absolute frequency, was the criterion selected for ranking the medical words classified applying this first procedure, because it provided the best return – i.e., the smallest number of word types to obtain the highest coverage results. In relation to the second procedure, it included medical words that only occurred in the medical corpus. Following the second procedure, the medical word types were ranked by their highest absolute frequency of occurrence in the medical corpus. A total of twenty-three 1,000 medical word lists were created, including words beyond the GLS, AWL, and EAP Science List.

The creation of a series of medical word lists, using the above mentioned twofold methodology (i.e., semantic rating scale and corpus comparison approach), has made possible the identification of the number of words (vocabulary load) required for students of medicine in general and for non-native medical students in particular to be able to cope with the lexical demands of medical textbooks. The enormous number of medical words to learn highlights the importance of acquiring subject-specific (medical) vocabulary as early as possible.

Conclusions and Implications

The use of the twofold methodology (i.e., semantic rating scale and corpus comparison approach) has enabled the creation of a comprehensive set of medical word lists to deal with the lexical demands of medical textbooks. The series of medical word lists here developed can serve several purposes. For instance, these medical word lists can be used as a guide for designing the vocabulary syllabus of an English for Medical Purposes course, making more informed decisions on the vocabulary worth focusing on when planning and teaching an ESP lesson, assessing and testing the learner's performance, and instructing medical students in the vocabulary learning strategies necessary for them to take control of the learning of content area (medical) vocabulary inside and outside the ESP classroom.

The text coverage results presented in this study demonstrate the large numbers of content area (medical) vocabulary – at least 26,000 different word types – making up medical texts. These words range from very high frequency words to many words occurring only once in the corpus, and represent an enormous amount of learning for both native speakers and non-native speakers training to be doctors. This very large number of medical words to learn stresses the importance of devising a plan for ESP reading courses, a plan that underlines the value of (1) strategy training in the ESP reading courses for medical students, (2) learning medical vocabulary as early as possible, (3) having a reasonable vocabulary size before starting medical study, and (4) testing the vocabulary size of ESP learners.

Some comments on the perceived limitations encountered during this investigation in relation to the size of the corpora, the identification of medical words, the nature of medical texts used for making the medical corpus, and the pedagogical value of the medical word lists created can be summarised as follows:

1. *Size of the corpora.* It is not always possible to adopt the corpus comparison approach with different corpora of similar or the same size, but comprising different (e.g., medical vs. general) topics, as in the case of the present investigation. A common solution to this problem is to normalize the frequency scores to a common base.
2. *Identification of medical vocabulary.* The replication of Chung and Nation's (2003) semantic rating scale involved the identification of thousands (at least 26,000) o f medical words occurring in the medical

corpus used for creating the lists. In spite of the usefulness of the semantic rating scale for making decisions on the number of content area vocabulary found in medical texts, its implementation proved to be a very demanding and time-consuming.

3. *Medical corpus limited to textbooks.* The medical texts included in the medical corpus compiled for the present investigation was restricted to textbooks. For future research to estimate the vocabulary load of medical texts, it would be worth including a variety of text types (such as medical articles in specialised journals and scientific magazines, book chapters, technical reports, and laboratory manuals) when creating a specialised corpus of medical texts written in English.
4. *Pedagogical value of the medical word lists.* The results of this investigation have shown that readers of medical textbooks need to know about 26,000 medical word types beyond existing word lists – as represented by the GSL, AWL, and EAP Science List, respectively – to be able to meet the lexical demands of medical textbooks. As detailed in Appendix A, the pedagogical value of the last two-thirds of the new medical word lists (i.e., around 16,000 medical word types needed for an additional 1% cumulative text coverage) is questionable. The acquisition of 26,000 medical words is a vocabulary learning goal that seems unrealistic to achieve in the restricted time span (one to two years at most) of most English of Medical Purposes reading courses. The need to learn these 26,000 medical word types clearly indicates that the technical vocabulary of medicine is very large and represents a major learning burden for the students learning to read medical texts written in English.

Vocabulary expansion of medical terms should be an important goal for teachers of English for Medical Purposes. In order to help ESP learners better cope with the lexical demands of medical texts and the large number of medical words required to achieve an adequate lexical threshold, ESP teachers need to:

1. Design a lexical syllabus to teach the vocabulary learning strategies, such as guessing from context, using mnemonic techniques, using word cards, doing extensive reading, that enable medical students to cope with most of the new vocabulary independently.
2. Encourage learners to do extensive reading on topics that address the vocabulary they are trying to learn.
3. Promote the use of genuine lexical contexts and provide authentic examples of medical vocabulary. Examples of authentic reading materials for meeting and learning medical terms in context are medical textbooks like those used to create the medical corpus mentioned in the present study.
4. Emphasise word relationships such as lexical bundles, word frequency, and phraseology.
5. Set ambitious vocabulary learning goals for your students of around 50 words per week.
6. Group the vocabulary that needs to be learnt in a manageable format (e.g., word family lists).

In conclusion, it is important to equip medical students in the ESP classes at university with the vocabulary learning strategies necessary to manage the acquisition of the massive number of words required to achieve good reading comprehension of medical texts written in English.

References

Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, 1–13. https://doi.org/10.1093/applin/amt018

Browne, C. (2013). The new general service list: Celebrating 60 years of vocabulary learning. *The Language Teacher*, *37*(4), 13–16.

Chen, Q., & Ge, G.-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*(4), 502–514.

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

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

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

Coxhead, A., & Hirsh, D. (2007). A pilot science word list for EAP. *Revue Francaise de Linguistique Appliqueé*, *7*(2), 65–78.

Coxhead, A., & Quero, B. (2015). Investigating a Science Vocabulary List in university medical textbooks. *TESOLANZ Journal*, *23*, 55–65.

Engels, L. K. (1968). The fallacy of word-counts. *IRAL - International Review of Applied Linguistics in Language Teaching*, *6*(3), 213–231. https://doi.org/10.1515/iral.1968.6.1-4.213

Fauci, A. S., Braunwald, E., Kasper, D. L., Hauser, S. L., Longo, D. L., Jameson, J. L., & Loscalzo, J. (2008). *Harrison's principles of internal medicine* (17th Edition). New York: McGraw-Hill. Retrieved from http://highered.mcgraw-hill.com/sites/0071466339/information_center_view0/table_of_contents.html

Fraser, S. (2005). The lexical characteristics of specialized texts. In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (Eds.), *JALT2004 conference proceedings* (pp. 318–327). Tokyo: JALT. Retrieved from http://jalt-publications.org/archive/proceedings/2004/E115.pdf

Fraser, S. (2006). The nature and role of specialized vocabulary: What do ESP teachers and learners need to know. *Hiroshima Studies in Language and Language Education*, *9*, 63–75.

Fraser, S. (2007). Providing ESP Learners with the Vocabulary They Need: Corpora and the Creation of Specialized Word Lists. *Hiroshima Studies in Language and Language Education*, *10*, 127–145.

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

Goldman, L., & Ausiello, D. (Eds.). (2008). *Cecil textbook of internal medicine* (23rd edition). Philadelphia, PA: W.B. Saunders Elsevier. Retrieved from http://www.us.elsevierhealth.com/cecil-medicine/goldman-cecil-medicine-expert-consult/9781416028055/

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *Range [Computer software]*. en, Wellington, New Zealand: Victoria University of Wellington.

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

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–30.

Hwang, K., & Nation, I. S. P. (1989). Reducing the vocabulary load and encouraging vocabulary learning through reading newspapers. *Reading in a Foreign Language*, *6*(1), 323–335.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, *41*(2), 235–253. https://doi.org/10.1002/j.1545-7249.2007.tb00058.x

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? *Special Language: From Humans Thinking to Thinking Machines*, 316–323.

Laufer, B. (1992). How much lexis is necessary for reading comprehension. In H. Béjoint & P. J. Arnaud (Eds.), *Vocabulary and applied linguistics* (Vol. 3, pp. 126–132). London: Macmillan.

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.

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. https://doi.org/10.1016/j.jeap.2016.01.008

Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes*, *28*(3), 183–198.

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes*, *25*(2), 235–256.

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

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, *63*(1), 59–82.

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

Nation, I. S. P. (2016). *Making and Using Word Lists for Language Learning and Testing*. Amsterdam: John Benjamins Publishing Company. Retrieved from http://www.jbe-platform.com/content/books/9789027266279

Quero, B. (2015). *Estimating the vocabulary size of L1 Spanish ESP learners and the vocabulary load of medical textbooks.* (Unpublished PhD thesis). Victoria University of Wellington, Wellington, New Zealand.

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

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies*, *7*(2), 105–125.

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

Schmitt, N., & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, available on CJO2012. https://doi.org/10.1017/S0261444812000018

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

Wang, K., & Nation, I. S. P. (2004). Word meaning in academic English: Homography in the Academic Word List. *Applied Linguistics*, *25*(3), 291–314. https://doi.org/10.1093/applin/25.3.291

Ward, J. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language*, *12*(2), 309–324.

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

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

Yang, M.-N. (2015). A nursing academic word list. *English for Specific Purposes*, *37*, 27–38. https://doi.org/10.1016/j.esp.2014.05.003

Appendix A

Text Coverage and Frequency of Occurrence of the Medical Corpus by the GSL, AWL, EAP Science List and the Twenty-Six Medical Word Lists

Word List	Tokens #	Tokens %	Types #	Types %	Families #
GSL1	3,021,029	55.62	3,291	5.95	981
GSL2	324,020	5.97	2,415	4.36	886
AWL	447,254	8.23	2,418	4.37	565
EAP Sc. List	329,236	6.06	1,288	2.33	316
MGEN1	461,169	8.49	1,000	1.81	n/a
MGEN2	98,853	1.82	1,000	1.81	n/a
MGEN3	47,476	0.87	1,000	1.81	n/a
MED1	280,114	5.16	1,000	1.81	n/a
MED2	79,208	1.46	1,000	1.81	n/a
MED3	44,413	0.82	1,000	1.81	n/a
MED4	29,254	0.54	1,000	1.81	n/a
MED5	21,085	0.39	1,000	1.81	n/a
MED6	16,127	0.30	1,000	1.81	n/a
MED7	12,593	0.23	1,000	1.81	n/a
MED8	10,018	0.18	1,000	1.81	n/a
MED9	8,168	0.15	1,000	1.81	n/a
MED10	6,635	0.12	1,000	1.81	n/a
MED11	5,546	0.10	1,000	1.81	n/a
MED12	4,773	0.09	1,000	1.81	n/a
MED13	4,000	0.07	1,000	1.81	n/a
MED14	3,502	0.06	1,000	1.81	n/a
MED15	3,000	0.06	1,000	1.81	n/a
MED16	2,978	0.05	1,000	1.81	n/a
MED17	2,000	0.04	1,000	1.81	n/a
MED18	2,000	0.04	1,000	1.81	n/a
MED19	2,000	0.04	1,000	1.81	n/a
MED20	2,000	0.04	1,000	1.81	n/a
MED21	1,333	0.02	1,000	1.81	n/a
MED22	1,000	0.02	1,000	1.81	n/a
MED23	1,000	0.02	1,000	1.81	n/a
Words outside the lists	159,956	2.94	19,942	36.03	0
Total	5,431,740	100.00	55,354	100.00	2,748

Acknowledgements
I would like to thank Emeritus Professor Paul Nation and Dr. Averil Coxhead of Victoria University of Wellington for their unfailing assistance and advice on an earlier version of this article. I am also grateful to the two anonymous *TESOL International Journal* reviewers for their constructive critiques and comments that have helped enhance the quality of this work.

About the Author
Betsy Quero is a language teacher and researcher with over fifteen years of experience. She has taught in New Zealand, England, and Venezuela. She holds a PhD in Applied Linguistics and is currently investigating the vocabulary load of academic texts. Her research interests include ESP pedagogy, specialised word lists, vocabulary testing, and language learning strategies.