Estimating the vocabulary size of L1 Spanish ESP learners and the vocabulary load of medical textbooks

by

Betsy Quero

A thesis submitted to Victoria University of Wellington in fulfilment of the requirements for the degree of Doctor of Philosophy in Applied Linguistics

Victoria University of Wellington

2015

Abstract

The main goal of this research on the vocabulary load of medical textbooks is to use a corpus-based approach to investigate the number of words learners of English for Specific Purposes (ESP) need to meet the lexical demands of medical textbooks written in English, and achieve a good reading comprehension of such texts. This thesis explores the relationship between vocabulary size of prospective learners of ESP in two Spanish medium universities in Venezuela, and the vocabulary load of medical textbooks.

The two main variables considered to calculate the vocabulary load of medical textbooks were the vocabulary sizes of ESP undergraduate students, and the lexical text coverage of medical textbooks. In order to determine the vocabulary size of ESP learners, a monolingual version of Nation and Beglar's (2007) Vocabulary Size Test (VST) and a Spanish bilingual version of the same test, which was translated for the purpose of the present study, were administered to a group of 408 English as a foreign language (EFL) learners who are speakers of Spanish as a first language (L1). For determining the lexical text coverage of medical textbooks, a written medical (med1) corpus of 5.4 million tokens, and a general comparison corpus of the same size were used. The lexical text coverage was estimated in two ways, with the help of existing word lists such as the General Service list (GSL; West, 1953), the Academic Word list (AWL; Coxhead, 2000), the Pilot Science List (PSL; Coxhead and Hirsh, 2007), and with the help of the British National Corpus/Corpus of Contemporary American English (BNC/COCA; Nation, 2012) word family lists. Additionally, medical words across the high, mid and low-frequency bands were identified using a two-level semantic rating scale designed for this purpose, and a set of over thirty 1,000 medical word lists developed. The new medical word lists created using the med1 corpus were then validated on a different medical corpus (i.e., the med2 corpus) of a similar size (5.8 million tokens).

The major finding of this thesis was that in order to reach the 98% lexical threshold required to achieve optimal reading comprehension of medical textbooks written in English, besides the top 3,000 most frequent word families of the English language, ESP learners doing medical studies would need to know at least 26,000 words related to the medical field. Moreover, this research found that when adding medical words from the existing word lists

(e.g., GSL, AWL, Pilot Science List, and BNC/COCA lists) to the new medical word lists created here, the overall coverage of medical words gets close to 37% of the running words in the medical text.

In relation to the vocabulary size of the EFL learners tested, this research determined: (1) the average vocabulary size of the Spanish as a first language (L1) ESP students taking the VST which was around 6,000 word families, and (2) the point at which it becomes more efficient for the medical students to focus on medical words rather than general words (e.g., knowledge of the 5,000 most frequent families). The results of this research placed particular emphasis on the medical words that ESP students meet in medical textbooks, but may not learn through existing general and academic word lists.

To my beloved family for reminding me of the important things in life.

Acknowledgements

I would like to express my sincere appreciation to many individuals and institutions for various types of help with this research project. First and foremost, my most heartfelt gratitude goes to my primary supervisor, Emeritus Professor Paul Nation, for his endless willingness to share his vast knowledge and expertise in the field of vocabulary studies. Without Paul's strong commitment, patient guidance, and insightful input, this thesis simply would not be. Likewise, I am grateful to my secondary supervisor, Dr Averil Coxhead, for her tireless support and helpful suggestions throughout this project. Averil's continuous encouragement and kindness will always be remembered. I could not honestly wish for a better supervisory team than Paul and Averil for the present research topic.

Besides my supervisors, my sincere appreciation also goes to Dr Stuart Webb, Dr Helen Basturkmen, and Dr Marlise Horst for taking time out from their busy schedules to serve as my examiners, and make useful suggestions and comments for the improvement of this thesis.

I also wish to thank the following institutions, colleagues, fellow students, advisers, consultants, proof readers, participants and individuals for their various forms of collaborative support and contributing their bit to the successful completion of this thesis:

- University of Los Andes (ULA, Mérida, Venezuela):
    - Department of Academic Staff Matters (DAP in Spanish).
    - Faculty of Humanities and Education.
    - Colleagues and administrative staff in the School of Modern Languages.
    - English for Medical Purposes adviser: Dr Françoise Salager-Meyer.

- Victoria University of Wellington (New Zealand):
    - Academic and administrative staff in the School of Linguistics and Applied Language Studies (LALS): Vocabulary Discussion Group, Thesis group, and Morphology group.
    - Current and former graduate students, from LALS and various graduate schools at Victoria University, who provided valuable comments and advice

on different aspects of this project: Brian Strong, Cherie Connor, Diego Navarro, Friederike Tegge, Irina Elgort, Juan Rada, Joseph Sorell, Ken Smith, Mario Alayón, Mark Toomer, Melissa Bryan, Michael Rodgers, Micky Vale, Myq Larson, Nguyen Thi Cam Le, Oliver Ballance, Sai Hui, Shaun Manning, Tastuya Nakata, Tatsuhiko Matsushita, Teresa Mihwa Chung, Thi Ngoc Yen Dang, TJ Boutorwick, Yosuke Sasao, and Zihan Yin, among many others.

- Subject librarian for Applied Linguistics: Tony Quinn.
- Academic and administrative staff in the Spanish and Latin American Studies Programme, School of Languages and Cultures.
- Faculty of Humanities and Social Sciences: Research Grant Committee.
- Faculty of Graduate Research: Scholarship Office.
- Language Learning Centre Technology Specialist: Edith Paillat.
- Student Learning Support Advisers: Dr Deborah Laurs, Laila Faisal, Kirsten Reid, Madeleine Collinge, Ann Pocock, Dennis Dawson and others.
- Victoria International Officer: Kerry Taplin.
- Student Financial Advisers: Megan Dow, Haley Collier, Sarah Prisk, Susana Sanele and others.
- Student Counselling Services: David Mason, Kelly Atherton and others.
- Student Financial Support and Advice Manager: Maria Goncalves-Rorke.

- Co-translator of the localised for Spanish version of the VST: Blanca Guzmán.
- Reviewer of the translation of the Spanish version of the VST: Dr Carolina Miranda.
- Proof readers of the localised for Spanish version of the VST: Dr Camino Rea Rizzo, Dr Daniela Díaz Guisado, Dr Miguel Arnedo-Gómez, Anayibi Loboa, Mario García, Juany Murphy, Pedro Saura, and Teresa Neches.
- Administrators of the VST: Alexandra Toro, Ana María Arellano, Ana María Maldonado, Duida Figueroa, Ingrid Contreras, Ligia Rivas, Marinés Asprino, Dr Olga Muñoz, Dr Orlando Suescún, Robert Márquez, and Dr Teadira Pérez.
- Primary developer of the website (VocabularySize.com) used for the online administration of the VST: Dr Myq Larson.
- Statistical consultant: Dr Dalice Sim.
- VST data analysis consultant: Dr Erwin La Cruz.

- The broader international community of linguistic scholars who kindly replied to my email messages and were willing to share unpublished papers and the results of their latest research findings: Dr Camino Rea Rizzo, Dr Françoise Salager-Meyer, Dr Irina Elgort, Dr Nguyen Thi Cam Le, Dr Nikolaos Konstantakis, Dr Nadežda Stojković, and Dr Ann Bertels, among others.
- A special note of appreciation warmly goes to more than 400 students participating in the piloting and administration of the VST in Venezuela.

Last but not least, I owe an enormous debt of gratitude to my family and a long list of friends in different parts of the world for all their help and various kinds of support all these years.

My appreciation is much deeper than the words above can show.

¡Un millón de gracias a todos!

Table of contents

List of tables

List of figures

List of appendices

Chapter 1: Introduction

My motivation in conducting this research comes from my background as an English for Specific Purposes (ESP) teacher. From a pedagogical point of view, I believe that it is important that English as a Foreign Language (EFL) and English as a Second language (ESL) teachers approach vocabulary instruction in ESP contexts in a way that focuses students' attention on the words that are worth learning, i.e., the words that will help learners achieve an appropriate reading comprehension of academic texts.

For years, I have heard my EFL students attributing their failure to understand written academic texts to the fact that they do not know enough words. This legitimate concern, particularly among English for Medical Purposes (EMP) students, regarding the relationship between subject-field vocabulary and reading comprehension, triggered my interest in the specialised vocabulary of medical texts written in English. After all, as Alderson (2000, p. 18) argues, "the ability to recognise words rapidly and accurately is a predictor of reading ability."

With these vocabulary teaching and learning beliefs in mind, I look at the relationship between the vocabulary size of prospective ESP learners and the number of words that this group of learners needs to know in order to be able to reach a 98% lexical threshold; the threshold required to meet the lexical demands of medical texts written in English for this investigation into the vocabulary load of medical texts. Throughout this thesis connections are made between the vocabulary size of ESP learners and the number of words needed for an appropriate lexical coverage (95%-98%) of medical textbooks. In particular, the present study focuses on English for Medical Purposes learners, as a sub-group of ESP learners, with highly specialised vocabulary learning needs from the early stages of their medical studies at university level.

This research is deeply rooted in the quest for lexical shortcuts and principled vocabulary teaching methods for accelerating the growth of academic and specialised vocabulary in

ESP. In many ways this project mirrors other vocabulary studies which have given special attention to principled approaches for identifying the words needed for an appropriate receptive comprehension of academic and specialised texts (Chung & Nation, 2003, 2004; Coxhead, 2000; Gardner & Davies, 2014) from various subject-specific fields, and for developing academic (Coxhead, 2000; Xue & Nation, 1984), scientific (Coxhead & Hirsh, 2007), and discipline-specific (Fraser, 2005, 2009; Konstantakis, 2007, 2010; Mudraya, 2006; Ward, 1999, 2009) word lists with a pedagogical purpose. I believe the research produced in this thesis makes a positive contribution to principled approaches to identifying discipline-specific words.

Chapter 2 is a review of the relevant literature, namely, the relationship between vocabulary size and the vocabulary load of written medical texts. The first part discusses estimates of the receptive vocabulary size of university students. The second part explores the use of instructional word lists (i.e., general, academic, scientific, and subject-specific lists). The most commonly used test-formats (i.e., yes/no test, matching, multiple-choice) in designing vocabulary size tests in an L2 (second language) are also discussed. The final part of Chapter 2 examines some of the classifications that have been used to define, characterise, and identify levels of vocabulary, degrees of technicality of specialised words, and the technical vocabulary of medical texts. Chapter 2 concludes by presenting the research questions for the present study.

Chapter 3 reports on the procedures and results of the administration of the Vocabulary Size Test (VST) developed by Nation and Beglar (2007) to a group of 408 Spanish as a first language (L1) ESP learners doing undergraduate studies at two Spanish medium higher education institutions in Venezuela. Firstly, this chapter justifies the rationale behind a Spanish bilingual version of the VST. Then it explains the principles guiding the translation into Spanish and the ordering of the items of the bilingual and monolingual VST. This chapter concludes by reflecting on whether the bilingual VST could be a better predictor of vocabulary size than the monolingual VST, and on the importance of calculating the vocabulary size of ESP undergraduates at the start of their introductory courses in ESP.

Chapter 4 investigates the vocabulary load of medical textbooks using a medical and a general corpus of the same size (5.4 million tokens each) compiled specifically for this study. This chapter outlines the traditional approach to identifying medical vocabulary in

English. This identification involved a corpus comparison of the lexical profile results of the medical and general written corpora mentioned above. The medical corpus used is representative of the medical lexis encountered in the English-language textbooks in basic science, health and medical disciplines that are compulsory for all medical students regardless of their fields of specialisation. The data analyses carried out in this chapter to investigate the vocabulary load of medical textbooks was conducted in the following order: (1) regularising the word families of the existing word lists used, namely, the General Service List (GSL; West, 1953), the Academic Word list (AWL; Coxhead, 2000) and the Pilot Science List (PSL; Coxhead & Hirsh, 2007); (2) developing a semantic rating scale for identifying words related to health and medicine, (3) describing the corpus compilation criteria for the medical and general corpora used; (4) explaining the steps undertaken for making new medical word lists; (5) calculating the lexical coverage of the existing word lists and the new medical word lists created for this study with Range (Heatley, Nation, & Coxhead, 2002); and (6) reflecting on the lexical profile results obtained from comparing shared words in the medical (med1) corpus and the general corpus compiled for the present study.

Chapter 5 continues with the estimation of the vocabulary load of medical textbooks using the same two corpora (i.e., medical and general corpora comprising 5.4 million tokens each) and the same software (i.e., Range) from the previous chapter. However, in this phase, both the medical and general corpora are run on a more recent set of existing word family lists, namely, the British National Corpus/Corpus of Contemporary American English lists (BNC/COCA lists) developed by Nation (2012) which are different from the ones used in Chapter 4. In the first part of this chapter, the lexical profile results of the tokens[1], word types[2] and word families[3] in the BNC/COCA lists are presented. Then, the frequency and

[1] A *token* is a unit of word counting that counts every occurrence of any word form as a different running word or token (Nation, 2013).
[2] A *word type* is a unit of word counting that counts every occurrence of the same word form as one more occurrence of the same word (Nation, 2013).
[3] A *word family* is a unit of word counting that counts different inflected and derived forms closely associated to the same base word as the same word family. A word family consists of a headword (i.e., base word), and its word family members (i,e., its corresponding inflected and closely related derived word forms) (Nation, 2013).

distribution of the medical words across the twenty-five 1,000 BNC/COCA word lists are explained. In the final part, the link between the vocabulary size of English for Medical Purposes students and the vocabulary load of medical textbooks at various frequency levels – across the high, mid, and low-frequency bands proposed by Schmitt and Schmitt (2012) – are discussed.

Chapter 6 validates the medical word lists created in Chapters 4 and 5. Before presenting the results of the validation of the medical words lists, the compilation process of a different medical corpus (i.e., the med2 corpus) with similar characteristics to the med1 corpus (i.e., the corpus originally created to make the word lists used in Chapters 4 and 5) is explained. Finally, the behaviour of the medical word lists on the med2 corpus is reported and the lexical text coverage results of the med1 and med2 corpora on the GSL, AWL, Pilot Science List, and BNC/COCA word family lists are compared.

Chapter 7 includes an additional discussion of the results presented in Chapters 3, 4, 5 and 6, and considers the key implications of the findings in relation to the two aspects of needs analysis (Hutchinson & Waters, 1987) in ESP: lacks (i.e., the vocabulary size of L1 Spanish ESP learners), and necessities (the vocabulary load of medical textbooks) reported in this thesis.

Chapter 8 offers concluding remarks. It discusses the minimum vocabulary size recommended for English for Medical Purposes students at the beginning of an English for Medical Purposes reading programme to be able to competently read medical textbooks. It also considers whether focusing first on the learning of the existing word lists (i.e., GSL, AWL, Pilot Science List, BNC/COCA lists) before learning the vocabulary of the more specialised medical word lists provides a more effective path to achieve the 98% lexical threshold. Pedagogical recommendations are provided on possible teaching principles that may be used to help English for Medical Purposes students who are not native speakers of English achieve the lexical threshold (98%) required for optimal reading comprehension of medical textbooks. This thesis concludes by reflecting on the limitations of this study and possible future research.

1.1 The ESP background of the study

In this section, the background on ESP that has served as the initial motivation for the present investigation is explained. Firstly, the rationale behind this study which explores the relationship between the vocabulary size of a group of postsecondary school ESP learners and the vocabulary load of medical textbooks is provided. Secondly, the researcher's ESP context, including the researcher's ESP teaching and learning situation in Venezuela, and the content and organisation of the ESP course are described. Finally, this section concludes by reflecting on the value of this research for a particular group of ESP learners and teachers.

1.1.1 The rationale of the study

The rationale for the present study originates from the constant need in ESP research to identify which are the general and subject-specific words (see section 2.1.3 of the literature review for references) worth focusing on to be able to reach the lexical threshold (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt, Jiang, & Grabe, 2011) needed for adequate reading comprehension of academic texts. With one particular ESP subfield (i.e. medicine) in mind, this investigation proposes a methodology to estimate the vocabulary load of medical texts written in English.

This study was conducted taking into consideration a specific undergraduate population of ESP students, those in ESP reading classes at two Spanish medium universities in Venezuela. These ESP courses in Venezuela focus predominantly on the development of the discipline-specific reading skills of the ESP learners. This is also a common ESP teaching and learning situation in other Spanish and Portuguese speaking higher education institutions in Latin America (see Salager-Meyer, Llopis de Segura, & Guerra Ramos, in press, for an overview of EAP/ESP in Latin America).

First of all, a representative population of prospective ESP students in Venezuela was selected to estimate the vocabulary size of L1 Spanish ESP learners (see Chapter 3). Subsequently, this research takes into account the vocabulary size results presented in Chapter 3 as the starting point to investigate the vocabulary load of one specific discipline: medicine (see Chapters 4, 5, 6). Among the wide variety of ESP reading programmes on

offer for Venezuelan undergraduates (e.g., ESP for the Humanities and Social Sciences, International Relations, Engineering, Medical and Health Sciences, Natural Sciences, Forestry, Information Technology, Political Sciences, and Statistics, among others), the present research narrows its focus in Chapters 4, 5, and 6 to specifically investigate the lexical demands of medical texts. The decision to select medicine as the specialised content area for the present investigation is based on the fact that, since the second half of the twentieth century, medical English has been cited more and more in the literature as the lingua franca (Csilla, 2009; Frînculescu, 2009; Maher, 1986; Salager-Meyer, 2014; Showell, Cummings, & Turner, 2010), par excellence, for international and intranational communication in the Western world.

1.1.2 The teaching and learning context of the study

English as a foreign language (EFL) is a compulsory subject in Venezuelan secondary schools for an average of five years. The current Venezuelan primary and secondary school curricula claim to be based on a holistic view of education (see Miller, 2000 for more details on the holistic approach in education) as explained in a report issued by the Venezuelan Ministry of Education in 2007 (see http://www.redalyc.org/articulo.oa?id=35603920 for more details of this report in Spanish). Nevertheless, in my experience as a secondary school teacher of English as a foreign language (EFL) in Venezuela, most of the class time in the EFL classroom is used presenting and practising the English language in an analytic way. That is, a large amount of the class time is devoted to explaining English grammatical rules and doing grammar and translation exercises with very little focus on communication or teaching language learning strategies.

At university level, there are English for General Purposes (EGP), English for Academic Purposes (EAP) and English for Specific Purposes (ESP) courses on offer. The ESP instruction at higher education institutions is a common subject for most university degrees in Latin America these days (Salager-Meyer et al., in press). ESP courses at university level are generally ESP reading courses. Particularly for the more science-oriented university degrees, ESP is a compulsory subject, while it is optional in other programmes. For those degrees where ESP courses are compulsory, the ESP instruction in Venezuela takes place for at least one year or two semesters. The second year is optional for most ESP programmes. The ESP courses for first and second year undergraduates in Venezuela have

an average of 35-40 students per class. Generally speaking, these classes have a total of four contact hours per week, and normally meet twice a week for two hours.

In the specific case of the ESP learners enrolled for an undergraduate degree in the Faculty of Medical and Health Sciences in Venezuela, many of them take their first ESP course as part of their bachelor degree (e.g., Bachelor of Health Sciences, Bachelor of Dental Surgery, Bachelor of Science in Human Nutrition, Bachelor of Science in Bioanalysis, and Bachelor of Pharmacy, among others), usually in the first year of their university studies. Most medical and health sciences degrees in Venezuela last between 5 and 7 years on average.

In general, L1 Spanish speakers in Venezuela are very unlikely to need English as a means of instruction or communication in their habitual health and medical settings; however, this particular group of ESP learners needs to comprehend specialised texts written in English to be able to keep up to date throughout their university degrees with the latest scientific breakthroughs which, in most cases, are first published in English.

In relation to the language of instruction in the ESP reading courses, it is Spanish, the L1 of the ESP learners and most ESP teachers in the Venezuelan context. Additionally, the classroom discussions, explanations and assessment are all conducted in Spanish. English is, however, the language of the selected readings for discussion in this ESP reading classes. Even though, the ESP teachers and students talk about the ESP written materials in their L1 (i.e., Spanish), the inclusion of examples in English in the class explanations is frequent. For instance, English is present in the examples included in the course handouts, on the information written on the board and on the slides presented by the ESP teacher. The purpose of including examples in English is to illustrate salient linguistic features of the specialised texts being described in the ESP classes.

1.1.3 The content of the ESP courses

The teaching of English for general purposes at Venezuelan higher education institutions aims at the development of the four language skills and the communicative competence of the EFL learners. The teaching of ESP uses a theme-based model of content-based

instruction (CBI) (see Brinton, Snow, & Wesche, 2003 for an overview of content-based instruction in L2 contexts).

The ESP reading courses that adopt a theme-based model of CBI in Venezuela are characterised by: (1) being content oriented rather than content focused, (2) taking into account the interests and needs of the ESP leaners, (3) negotiating with the students the content area/discipline-specific topics to be discussed in the ESP reading classes, (4) providing the ESP learners with opportunities to gain new knowledge on content area topics of their interest, and (5) being taught by language specialists. This is achieved by including content learning opportunities that are linked, but independent of the content objectives of the mainstream discipline-specific subjects, (6) reinforcing the knowledge of disciplinary content that the ESP learners have been simultaneously acquiring in the L1 discipline-specific subjects, and (7) developing in the ESP learners the reading skills required for appropriate receptive written comprehension of their discipline-specific texts.

The topics to be included in the ESP courses in Venezuela are negotiated with the ESP learners at the start of the ESP course and include both discipline-specific topics being taught in the discipline-specific courses they are taking as part of the ESP learner's undergraduate training, as well as other discipline-specific topics the students are interested in reading. Typically, the main topics included in the course outlines of the discipline-specific courses taken in the L1 of the ESP students are summarised in a list. Then, the ESP learners are asked to select from this list of discipline-specific topics the topics they would like to discuss in the ESP readings courses.

Likewise, the ESP teacher is constantly asking for cooperation from the content specialists/lecturers of the discipline-specific courses. For instance, the ESP teacher is frequently consulting with content specialists on a variety of issues such as: (1) the current and future EFL learning needs of the ESP learners in their particular specialised subject areas; (2) relevant readings (e.g., mainly from authentic reading materials such as extracts from specialised reference books, textbooks, and periodicals) for the ESP reading course, and (3) the adequacy of the discipline-specific topics for inclusion in the ESP classes. It is also important for the ESP teacher to continue consulting with the content specialists at different stages of the ESP course. For example, in the planning stage when deciding on the linguistic features and topics to be included in the ESP course, the ESP teacher can

interview content specialists and ask them about the most pressing L2 learning needs of prospective ESP learners, as well as about timely specialised topics for discussion in the ESP classes. Additionally, in the middle of the ESP course the ESP teacher can consult again with the content specialists as a way of reassessing the needs analysis conducted in the planning stage of the course. Also, as a way of continuously monitoring the relevance of the teaching approach and language needs of the ESP learners, follow up meetings between the ESP and content specialists can be held.

For instance, in the particular case of the ESP teacher working with learners from a specific subfield of ESP such as English for Medical Purposes, it is common for the English for Medical Purposes teacher (i.e., the language specialist) to consult with the medical teachers/lecturers (i.e., the content specialists) of the discipline-specific subjects (i.e., medical and health sciences subject areas) the medical students are simultaneously studying in their L1 (i.e., Spanish, in the case of the medical undergraduates mentioned in the present study) and L2 (i.e., English for Medical Purposes).

The ESP teachers not only need to be familiar with the specific linguistic features of the subject area of specialisation of the ESP learners, but also need to have, at least, some general understanding of the main discipline-specific topics being discussed in the ESP lessons. Nevertheless, it can be foreseen that in a relatively short time span, of less than two years, most ESP learners may have developed a deeper understanding of a wider range of discipline-specific topics than the ESP teacher may have. In this respect, it is important for the ESP teacher to be constantly looking for ways to benefit from the more in-depth content area knowledge the ESP learners are capable of bringing into the ESP classroom. As a way of making the most of this situation, it is a common practise in the ESP classes to promote group discussions, peer work, and short presentations by the ESP learners. ESP students, in this particular context, play an active role and feel motivated to participate in a cooperative reading environment that fosters learning and exchange of ideas.

Likewise, it is important that throughout the whole teaching/learning process the ESP teacher is confident with the specialised topics selected, willing to promote discussions about relevant specialised topics that are also within his/her understanding, and focused on the teaching of the L2 learning strategies conducive to developing independent and competent L2 readers in their particular specialised content areas.

From a linguistic perspective, three main content selection criteria are considered when choosing the reading materials for the ESP courses in Venezuela, namely language, subject matter, and level.

1. *Language:* The readings should include the linguistic features (i.e., morphological, syntactic, semantic, pragmatic) that characterise specialised texts from the specific subject areas of specialisation of the ESP learners.
2. *Subject matter*: the topics of the readings are selected after considering the results of the needs analysis conducted with the subject matter experts (i.e., lecturers who are content specialists), the ESP learners, and other ESP teachers. For instance, in the case of English for Medical Purposes, the medical teachers are consulted on the most appropriate content and skills to support the medical study of the learners. In addition, discipline-specific topics can be selected from the content of course outlines of relevant discipline-specific courses the ESP students are taking. By recycling content already introduced in the learners' L1 and/or gaining new information on familiar specialised topics, the ESP course helps support the students' subject area of study.
3. *Level:* the linguistic features and subject matter of the reading materials should be at an adequate linguistic level to engage the students in the class discussions, and foster L2 learning and comprehension of the discipline-specific texts.

1.1.4 The value of this research for this particular group of ESP learners

The vocabulary size component of this research (see Chapter 3) is aimed at estimating the general vocabulary size of postsecondary ESP learners enrolled at two Spanish medium higher education institutions in Venezuela. One of the main reasons for choosing the Schools of Medicine, Engineering, Humanities and Social Sciences, Natural Sciences, and Farming Sciences for administering the VST, is related to the fact that these were the schools where there were colleagues willing to collaborate in the administration of the VST to their undergraduate students. One important feature all these test takers had in common was that they were only in the first couple of weeks of their first ESP course at university, which in most cases was their first semester at university, as well. This also means that the discipline-specific knowledge this population of ESP programmes have in their L1 was based on the content area knowledge they had acquired in their secondary school education.

That is, the vocabulary size of their general English was measured based on the level of English they gained during the five years of EFL in secondary school. Since all the ESP learners who sat the VST in Venezuela, regardless of their undergraduate degree, were being assessed on the number of general English words they had learnt during their secondary school years, measuring the vocabulary size of a wide range of first year ESP undergraduates would also provide a reliable estimate of the vocabulary size of first year medical students at the start of their first ESP course.

The three medical vocabulary studies that follow (see Chapter 4, 5, and 6) are specifically aimed at postsecondary English for Medical Purposes students about to start their first English for Medical Purposes course at university level. As noted above, the level of discipline-specific knowledge of the group of undergraduate medical students who took the VST was based on the natural sciences subjects they did at secondary school, (such as biology, chemistry, etc.). That is, the disciplinary knowledge of this particular group of medical students is very similar to the other groups of ESP learners who sat the VST in Venezuela and whose test results are described in Chapter 3 of the present investigation.

In the researcher's view, ESP teachers in Venezuela face two main challenges when teaching ESP reading courses. The first challenge has to do with the different teaching approaches in secondary schools versus higher education institutions. As mentioned above, while in Venezuelan secondary schools an analytical approach (i.e., including a variety of grammar translation exercises) is common, in higher education institutions a theme-based model of CBI is more common. A way for ESP teachers to narrow this gap – between EFL teaching approaches in secondary schools versus higher education institutions in Venezuela – is by training ESP students in the appropriate L2 learning strategies required to become proficient readers and to achieve adequate reading comprehension of specialised texts from their specific subject areas. The other challenge faced by the ESP teacher is related to the fact that Spanish speakers, in general, are more willing to communicate orally. The preference for oral exchanges of information among this particular group of ESP learners in Venezuela makes it more challenging for the ESP teacher to engage the students in ESP readings activities that lack some form of oral communication (e.g., group work with peer support). This second challenge can be overcome by promoting oral activities (such as group work, class discussions, mini talks, project work, and presentations, among others) as the preferred form of communication in the ESP classroom.

One of the major advantages of this particular group of ESP learners in Venezuela relates to the fact that they are L1 speakers of a Romance language, Spanish, and that both languages (i.e., English and Spanish) share a large proportion of Graeco-Latin cognates and loan words. For instance, as reported in the vocabulary size chapter of this thesis, around 54.29% of the target or tested words in the VST are English/Spanish cognates (see Chapter 3). Another strength of this group of ESP students lies in the fact that they have been exposed to English as a foreign language as a compulsory subject in secondary school for at least five years.

By first estimating the number of words that a group of postsecondary school ESP learners know at the start of their first ESP course at university, then calculating the number of words in medical textbooks required to achieve the 98% lexical threshold, and finally investigating the relationship between the vocabulary size of ESP learners and the vocabulary load of medical texts, the present research provides ESP teachers with valuable information to make more informed decisions on the lexical demands of English for Medical Purposes reading courses. That means that when ESP teachers embark on the course planning and design, teaching, and assessment of their English for Medical Purposes learners, the ESP teachers are able to match the lexical demands of medical textbooks closer to the lexical needs of their specific group of learners. For example, by pointing out to ESP teachers and course designers which are the words in medical texts worth focusing on, this research can contribute to reduce the learning burden of vocabulary in specific specialised subject areas like medicine.

Another example of the value of the present research is related to the estimation of the vocabulary size of prospective ESP learners and the creation of subject-specific lists. That is, once the vocabulary size of a particular group of ESP learners has been estimated (an average of 6,000 word families which is the result of the study reported in Chapter 3), then this research provides medical word lists including frequent medical words that can serve as the starting point for new vocabulary that needs to be introduced, practiced and recycled. Also, the words in English this group of ESP learners already knows can easily be retrieved in the vocabulary activities linked to the specialised L2 readings.

To sum up, one of the main goals of this research is to propose a methodology for the creation of discipline-specific word lists (i.e., medical word lists in the particular case of

the present study) that include the most salient vocabulary in medical texts. After explaining the methodology and presenting these discipline 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-based approach.
2. Explore the relationship between the vocabulary size and the vocabulary load of a group of ESP learners.
3. Lessen the vocabulary learning load when acquiring the discipline-specific lexis required for achieving an appropriate reading comprehension of medical texts.
4. Provide guidelines for the creation of medical word lists organised by levels of frequency and salience (see Chapters 4, 5, 6, 7).
5. Suggest the use of the word lists developed here as a guide for supporting the learning of specialised vocabulary in English as a foreign language by undergraduate medical students.

Chapter 2: Literature review

2.1 How large a vocabulary do university students need?

University students need to know or need to learn large numbers of words to be able to meet the lexical demands of discipline-specific texts in order to gain good reading comprehension of academic texts at university level. Word knowledge is a key component of language acquisition, and without enough receptive vocabulary, it is not possible to achieve good reading comprehension of academic or specialised texts (Laufer, 1989; Nation, 2013).

The number of English words that native and non-native university students know has been strongly linked to the acquisition of competence in reading. Second language (L2) vocabulary researchers (Nation, 2006; Schmitt, 2008) have concluded that a vocabulary of as many as 8,000-9,000 word families is necessary for general reading comprehension of authentic written texts (e.g., novels or newspapers). Such estimates on the receptive vocabulary size of L2 learners are based on the British National Corpus (BNC; Nation, 2005) lists and a lexical coverage figure of 98%. Milton and Treffers-Daller (2013, p. 152) affirm that "it is entirely possible that undergraduate students experience difficulties in reading because they do not know a sufficient number of words to understand what they read … It's students' vocabulary size which may explain, at least in part, why they are struggling to read academic texts."

The discussion of the number of words needed by readers of academic texts in this chapter includes three sections. First, the concepts of disciplinary language and disciplinary vocabulary are discussed. Second, the various linguistic aspects involved in knowing a word receptively are described. Third, the appropriate lexical thresholds of written academic texts are reported. Finally, the role of existing general, academic, scientific, and discipline-specific word lists is examined.

2.2 What does it mean to learn disciplinary language?

Terms such as 'varieties of language', 'dialects' and 'registers' are labels frequently attributed to various types of linguistic variation in English. For Millward and Hayes (2012, p. 345), "a dialect is a variety of a language distinguished from other varieties in such aspects as pronunciation, grammar, lexicon, and semantics." For Basturkmen (2006, p. 15), the term 'variety' when referring to 'varieties of language', "refers to registers of a language use, such as English in banking, English in medicine, English in academic settings, and everyday conversation."

To learn disciplinary language means to become familiar with a language variety (a dialect or a register) characteristic of a specific discipline or content area. More precisely, Millward and Hayes (2012, p. 352) define as occupational dialects a group of discipline-specific texts (medical, engineering, legal, philosophical, cooking, tourism, and business, among others) that share a set of common linguistic features and are representative of a given language variety. A particular language variety (i.e., occupational dialect or register) can be distinguished from other language varieties by distinctive linguistic features. These salient linguistic features are common to a range of topics within a specific discipline at various linguistic levels (e.g., phonetic and phonological, morphological, lexical, syntactic, semantic, and pragmatic).

Hyland (2002) calls for disciplinary specificity in the acquisition of the linguistic conventions of a particular language variety at all stages and proficiency levels. For Hyland, central attention should be given to the teaching and learning of disciplinary language in ESP courses at higher education institutions. Hyland (2009, p. 10) views language as "intimately related to the different epistemological frameworks of the disciplines and inseparable from how they understand the world." In this respect, he claims that the acquisition of appropriate disciplinary knowledge is closely linked to the disciplinary practices the ESP learners engage with others in their disciplines. At the pragmatic level, Hyland (2009, 2013) considers the constant interaction between context and meaning fundamental to promote adequate acquisition of disciplinary language.

The unavoidable encounter of all EAP learners with "specialist disciplinary language" (Woodward-Kron, 2008, p. 235), the complexity of disciplinary specificity in their

university studies, and the constant interplay between meaning, context, technicality and concrete versus abstract dimensions of words in disciplinary languages are aspects addressed by Woodward-Kron. In this respect, Woodward-Kron (2008, p. 227) states that "the distinct specialist language of the disciplines and their different areas of difficulty are aspects of learning specialist knowledge which students need to come to terms with if they are to be successful in their studies."

At the lexical level, Becher (1987) defines disciplinary terms (i.e., vocabulary) as those words that play a crucial role in identifying the key concepts in a given discipline, and that do not play the same central communicative role in other disciplines. For example, medical practitioners are likely to use medical vocabulary to refer to prevention, diagnosis, and treatment of diseases. In the case of engineers, for instance, they need language to talk about the design, building and use of structures and systems. These two disciplines use enough linguistic variation to allow, most of the time, for clear linguistic distinctions from one to another, and very specific linguistic features to communicate central concepts or ideas in each of these subject fields.

With the primary purpose of investigating the lexical demands and specificities of the disciplinary language of one particular discipline or content area (i.e., medicine), the present thesis looks at the vocabulary load of medical texts written in English.

2.2.1 What does it mean to know a word receptively?

As Read (2000, p. 25) states, the notion of what is meant by knowing a word "highlights the complex nature of vocabulary learning, which involves a great deal more than just memorising the meaning of a word." Knowing words in another language is therefore a complex notion. Several writers (Laufer, 1990, 1997; Nation, 2013; Richards, 1976) have expressed their views about the various aspects involved in knowing a word. For Richards (1976), knowing a word means knowing its frequency, register, syntactic features, and associations with other words, meaning, and semantic features. According to Laufer (1997, p. 141), knowing a word implies knowing its form (pronunciation and spelling), its morphology, its syntactic patterns, its meaning, its lexical relations, and its common collocations. For Nation (2013, p. 44), knowing a word involves knowing its form (spoken, written, and word parts), meaning (form and meaning, concept and referents, associations),

and use (grammatical functions, collocations, constraints of use) both receptively and productively. In this respect, Nation (2013) provides a more comprehensive account of what knowing a word means. He makes a detailed distinction between receptive and productive language knowledge of the various aspects involved in knowing a word.

Knowing a word receptively does not necessarily mean that we will be able to produce it. Recognising a word is not the same as producing it. That is, the difficulty of a word varies depending on whether it is going to be used receptively or productively. Nation (1990, p. 48) notes that it is easier to learn to recognise a word form and recall its meaning than it is to learn to produce the word at suitable times. Rough estimates from comparing the results of receptive and productive tests of the same words indicate that learning a word productively is 50 to 100 percent more difficult than learning it receptively (Morgan & Oberdeck, 1930; Stoddard, 1929; Waring, 1997).

This study focuses on receptive word knowledge by looking at the vocabulary size in a group of ESP learners, and the number of words needed by native and non-native university students to be able to meet the lexical demands of medical texts written in English. This literature review does not refer to productive word knowledge, because it is outside the scope of the present investigation.

2.2.2 Lexical threshold and the vocabulary load of academic texts

The importance of identifying the percentage of known words (vocabulary load) needed for unassisted reading comprehension has been investigated in several studies (Coxhead, Stevens, & Tinkle, 2010; Dang & Webb, 2014; Hirsh & Nation, 1992; Hu & Nation, 2000; Laufer, 1989; Nation, 2006; Nation & Waring, 1997; Webb & Nation, 2008). The first investigations (Laufer, 1989, 1992) on the lexical coverage of academic texts suggested a reading comprehension threshold of 95%. More recent research on the vocabulary load of academic written texts (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt et al., 2011) has moved the lexical threshold upwards and indicated that a higher lexical threshold of 98% or more is required for optimal unassisted reading comprehension. Chapters 3, 4, and 5 of this thesis investigate the vocabulary load necessary for reaching this optimal coverage of medical texts. That is, these three chapters explore

the number of words in medical texts required to be known to achieve a 98% lexical text coverage.

Additionally, in order to measure the vocabulary load of specialised texts a set of pedagogical word lists, that will be mentioned in the next section of this review, has been developed.

2.2.3 Pedagogical word lists in ESP

Word lists that have been created from various specialised corpora, using the criteria of frequency of occurrence, range, dispersion, lexical coverage, and ranked by frequency levels are described below. They have been developed primarily with an instructional purpose in mind. That is, these vocabulary lists have been made to inform language course writers, language test writers, material developers, curriculum designers, and language teachers on the most appropriate lexis to be taught, assessed, and learnt in ESP.

In this section, the impact of some well-known and extensively used general, academic and more specialised word lists in ESP is discussed. Specifically, research studies are discussed which have used some well-known general and academic word lists that have estimated the lexical coverage and vocabulary load of specific subject areas in ESP.

2.2.3.1 General word lists

Michael West's (1953) General Service List of English Words (GSL) is a high-frequency list that contains 2,000 word families (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. West's (1953) list was compiled using the criteria of frequency, ease or difficulty of learning, necessity, coverage, stylistic level, and intensive and emotional words. In the last 60 years, the GSL has been the most widely used high-frequency list for ESL and EFL curriculum planning, materials development and vocabulary instruction.

The GSL has been criticised for its age and lack of balance in some semantic fields (Hyland & Tse, 2007; Read, 2000, 2007; Richards, 1974), 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 more than two decades, vocabulary researchers constantly stated that the GSL was in need of revision or replacement (Coxhead, 2000; Hwang & Nation, 1989; Wang & Nation, 2004); however, it was not until its 60th anniversary that two 'new general vocabulary lists' were created (Brezina & Gablasova, 2013; Browne, 2013).

Nevertheless, in order to allow comparisons with previous studies that have looked at the number of general words in medical written texts, West's (1953) GSL is the general word list used in Chapter 4 to replicate the corpus-based approach. This approach has been traditionally used in the last 60 years to determine the number of general words in medical written texts in English.

Another series of general word lists that has been compiled recently by Paul Nation (2005, 2012) are the fourteen 1,000 BNC lists (Nation, 2005) and the twenty five 1,000 BNC/COCA lists (Nation, 2012). The words in these two sets of word family lists are ranked into lists based largely on the criteria of frequency and range. The twenty-five 1,000 lists extend the fourteen 1,000 lists and make use of data from the Corpus of Contemporary American English (COCA; Davies, 2014). The BNC/COCA lists (Nation, 2012) currently include a total of twenty nine 1,000 word family lists. The first twenty five 1,000 lists are lists of word families. The remaining four lists are lists of proper nouns, marginal words, transparent compounds, and abbreviations.

The BNC/COCA lists, and the previous set of general word lists above mentioned (i.e., West's first 1,000 GSL, and second 1,000 GSL) are the general word lists used in Chapters 4, 5, and 6 of this thesis. All these lists are available for free download at http://www.victoria.ac.nz/lals/about/staff/paul-nation as accompaniments of the Range software (Heatley et al., 2002).

2.2.3.2 Academic word lists

A number of word lists created in the past three decades have focused on the inclusion of a set of frequent words common to a wide range of academic texts. First, Xue and Nation (1984) created the University Word List (UWL) which resulted from combining words from four separate studies of academic vocabulary, namely, Campion and Elley's (1971) list, Praninskas' (1972) list; Lynn's (1973) list, and Ghadessy's (1979) list. Originally, as

reported by Xue and Nation (1984, p. 216), the UWL had 737 word families. Later, 99 new word families were added from the Barnard and Brown list (Nation, 1984). As a result, the UWL comprises now 836 word families that are not included in the first 2,000 GSL. That is, the UWL was designed to complement the GSL. For this reason, the UWL contains the academic vocabulary most common to different academic disciplines, and excludes the words in the GSL.

Let us now, in Table 2.1, look at the lexical coverage results of two investigations on the lexical profile of two different specialised subject-fields (i.e., Economics by Sutarsyah, Nation, & Kennedy, 1994; and Engineering by Ward, 1999). The coverage results of these two studies were obtained after running the specialised corpora they compiled, namely, an Economics corpus (Sutarsyah et al., 1994) and an Engineering corpus (Ward, 1999) through the GSL and UWL.

Sutarsyah, Nation and Kennedy (1994) compiled an Economics corpus of 295,294 tokens and a general academic corpus of a similar size (311,768 tokens) to investigate the vocabulary load of Economics texts. The results of the corpus comparison approach used for their investigation highlighted the narrowly-focused nature of specialised vocabulary from a particular subject area like Economics. Although both corpora (i.e., the Economics corpus and the general academic corpus) had a similar length, the Economics corpus contained half the number of word types (9,469) compared to the total number of word types (21,399) of the general academic corpus. Ward (1999) describes the design of a 2,000 high-frequency word list aimed at engineering students with low levels of English. Ward's (1999) list includes words frequently encountered in most engineering disciplines. For his study, Ward (1999) developed a corpus including 1,000,000 tokens of first year engineering textbooks. The texts chosen for this one million engineering corpus were from five engineering sub-fields, i.e., thermodynamics, mechanics, fluid mechanics, statistics and probability, and mechanics of materials. Ward's (1999) engineering list provides engineering students, in the first part of their degrees, with practical information on the words they need to focus on to achieve up to 95% coverage of engineering texts. That is, these lists were designed as an attempt to provide engineering students with a shortcut for the lexical demands posed by engineering texts. For Ward (1999) learning the GSL and the UWL is not really the fastest and most direct way to meet the vocabulary of engineering texts.

Table 2.1 Lexical coverage (%) by the GSL, and UWL on specialised texts

Researchers	Year	Corpus type	Number of tokens	GSL %	UWL %	Total %
Sutarsyah, Nation & Kennedy	1994	Economics textbook	295294	82.5	8.74	91.24
Ward	1999	Engineering textbooks	1000000	79.8	10	89.80

The coverage results in Table 2.1 indicate that 91.24% was the highest lexical coverage reported by research studies in the 1990s (Sutarsyah et al., 1994; Ward, 1999) calculating the percentage of words covered by the first and second 1,000 GSL, and the UWL on specialised corpora from two subject areas: Economics (Sutarsyah et al., 1994), and Engineering (Ward, 1999). In both cases, the total coverage by the GSL, and the UWL is lower than the recommended lexical threshold (i.e., 95%-98%) for good reading comprehension. These coverage figures however do not include proper nouns, marginal words, transparent compounds or abbreviations, which can account for around 5% of coverage.

More than a decade later, the UWL was replaced by Coxhead's (2000) Academic Word List (AWL). Like the UWL, the AWL is linked to the GSL. Since its creation, the AWL has been extensively used to learn, teach, and research academic vocabulary. According to Murphy and Kandil (2004, p. 64), the AWL is "the first listing of academic words that was developed through techniques of principled corpus analysis." To make the AWL, Coxhead (2000) gathered a corpus of 3,513,330 tokens. This corpus was comprised of a wide variety of academic texts from 28 academic subject areas, seven of which were grouped into one of the following four disciplines: (1) Arts, (2) Commerce, (3) Law, and (4) Science. The AWL contains 570 word families and covers around 10% of the tokens. For validating the AWL, Coxhead (2000) created a second academic corpus (comprising 678,000 tokens) which accounted for 8.5% coverage, and a non-academic corpus (i.e., a fiction corpus including 3,763,733 tokens) which covered 1.4% of the AWL. These coverage results of the AWL over the academic and fiction corpora confirm the academic nature of the words in the AWL. Its 570 word families are divided into 10 sublists. The word families for each sublist were selected based on the following three criteria: (1) *specialised occurrence*: choosing only words beyond the first 2,000 GSL, (2) *range:* including word family members occurring at least ten times in each of the four disciplines and in at least 15 of the 28 academic subject areas, and (3) *frequency:* selecting the word family members occurring

at least 100 times in the academic corpus which Coxhead (2000, p. 221) collected and analysed. As Coxhead (2000, p. 213) says, "the AWL shows learners with academic goals which words are most worth studying." Since the words included in the AWL occur more frequently in academic texts than in any other kind of text, these are words that L2 students of English for Academic Purposes (EAP) should know well. In this respect, Nation (2013, p. 301) points out that these words should be treated like high-frequency words in the case of L2 learners of EAP. That is, if a learner studies the GSL in conjunction with the UWL or later the AWL, there will be no repetition of word types found. (See Chapter 4 for further evaluation of the AWL).

Coxhead (2000, p.229) acknowledges the lack of a fully balanced comparison between the two academic corpora compiled to create and validate the AWL. According to Coxhead (2000), this imbalance is caused by the smaller size of the second academic corpus (i.e., 678,000 tokens) when compared to the 3,513,300 tokens of the academic corpus, and the difference in this academic comparison corpus between the number of tokens of each of the four disciplines (i.e., 82,000 tokens in Arts, 53,000 tokens in Commerce, 143,000 tokens in Law, and 4000,000 tokens in Science). Likewise, Coxhead's research (2000) on academic words has been criticised for being directly affected by the inclusion and exclusion criteria used to create the GSL (Nation & Webb, 2011), not including engineering (Ward, 2009) and medical texts (Chen & Ge, 2007), not providing frequency information for each of the members of the 570 word families (Paquot, 2007), not differentiating between the meanings and parts of speech (Paquot, 2007); and only basing the AWL on single word units (Durrant, 2009; Paquot, 2007). Also Paquot (2007) has criticised the AWL for using non-occurrence in the GSL as a criterion. Paquot (2007) considers that some of the most common 2,000 words of the English language may be words frequently used in some academic domains with a different meaning, although Wang and Nation's (2004) research suggests that this is not likely to involve many words. Moreover, the usefulness of a general academic vocabulary list like the AWL in subject-specific courses has been questioned in different corpus studies on academic vocabulary (Chen & Ge, 2007; Hyland & Tse, 2007; Paquot, 2007). In these studies, the authors argue that each single subject area poses different lexical demands for ESP learners, and these authors suggest that subject-specific word lists will cater better for the particular vocabulary needs of ESP learners.

Despite the criticism that the AWL has received so far, many other corpus-based studies (Chen & Ge, 2007; Cobb & Horst, 2004; Coxhead & Hirsh, 2007; Hyland & Tse, 2007; Martínez, Beck, & Panza, 2009; Vongpumivitch, Huang, & Chang, 2009) have consistently shown that, as Coxhead (2011) claimed, the AWL has a coverage of at least 10% or more in academic texts from different academic disciplines. Another study (Wang & Nation, 2004) that investigated word meaning found that only around 10% of its 570 word families contain homographs, and taking account of these homographs would exclude only three word families from the list – *intelligence, panel,* and *offset*. All these studies have shown why the AWL has been for more than a decade the most widely used list for academic vocabulary instruction in ESP and ESL settings.

A new Academic Vocabulary List (AVL) was created by Gardner and Davies (2014). The AVL was developed from a 120-million-token academic corpus of written English which is a subsection of the 425-million-token Corpus of Contemporary English (COCA; cited in Gardner & Davies, 2014). The AVL is based on 13,000 academic texts (i.e., academic journal articles, and academically-oriented magazine articles) from nine academic subject areas: (1) History, (2) Social Science, (3) Science and Technology, (4) Education, (5) Humanities, (6) Medicine and Health, (7) Law and Political Science, (8) Philosophy, Religion and Psychology, and (9) Business and Finance. The coverage of the AVL over the academic sections of the COCA (120 million tokens) and the BNC (33 million tokens) is 14% in both cases (Gardner & Davies, 2014). The words in the AVL are grouped in separate lists using three different units of counting, namely, word types, lemmas[4], and word families. Currently these different versions of the AVL are available for free download at http://www.academicvocabulary.info/download.asp in three formats (i.e. a 1991 word family AVL, a 3,015 lemma AVL, and a 20,845 word type AVL). Examples of core academic words from the AVL are *analysis*, *factor*, *figure*, *report*, and *value*. The AVL appeared too

[4] A *lemma* is a unit of word counting that counts the different inflected forms of a base word as the same lemma. Lemmas consist of a headword (i.e., base word), and its corresponding inflected word forms (Nation, 2013).

recently to be used in the present study. (See section 2.4 for further evaluation of Gardner & Davies, 2014).

The decision to use West's (1953) GSL and Coxhead's (1998) AWL in this study to estimate the lexical coverage of general and academic words in medical texts is based on the fact that these are two well-known word lists that have been traditionally used by ESP researchers to calculate the lexical demands posed by written academic texts. Moreover, the present study was carried out before new general (Brezina & Gablasova, 2013; Browne, 2013) and academic (Gardner & Davies, 2014) word lists were developed.

In relation to the lexical profile of spoken academic English, Dang and Webb (2014) looked at the vocabulary load of academic spoken English in a corpus of 1,691,997 tokens comprising four sub-corpora of the following specialised subject fields, namely, Arts and Humanities, Life and Medical Sciences, Physical Sciences, and Social Sciences. The results of their investigation indicated that an average vocabulary size of 4,000 and 8,000 word families plus proper nouns and marginal words provided a 96.05% and a 98% lexical coverage of academic spoken English, respectively. In the specific case of the coverage results of the Life and Medical Sciences sub-corpora of Dang and Webb's (2014) study, a 95.46% coverage was reached with 4,000 word families, and 98.05% with 13,000 word families. Of the four subject fields examined in Dang and Webb (2014), Life and Medical Sciences is the most demanding one in terms of lexical coverage. That means that in order to reach a 95% and 98% coverage of the Life and Medical Sciences sub-corpus, learners need a vocabulary size of 4,000 and 13,000 word families, respectively. The coverage of Dang and Webb's (2014) investigation over spoken medical texts agrees with the results of the present study (see Chapters 4, 5, and 6) in relation to the large number of words in medical texts needed to achieve the 98% optimal lexical threshold.

In the last decade, as summarised in Table 2.2, more recent vocabulary studies on the specialised texts from various subject fields (i.e., Business, Finance, Engineering, Linguistics, Sciences, and Medical and Health Sciences) have also calculated the lexical coverage of general and academic words using the GSL and the AWL (i.e., the UWL successor).

Table 2.2 Lexical coverage (%) of GSL and AWL over specialised texts organised by subject field

Researchers	Year	Corpus	Text type	Number of tokens	GSL % of tokens	AWL % of tokens	Total % of tokens
Cobb & Horst	2004	Learned section Brown corpus	Research articles	14283	78.53	11.60	90.13
Coxhead & Hirsh	2007	Science	Research articles & textbooks	1761380	71.52	8.96	80.48
Hyland & Tse	2007	Science, engineering & social sciences	Research articles	3213477	74	10.60	84.60
Hyland & Tse	2007	Science (subcorpora)	Research articles	838926	69	9.30	78.30
Martinez, Beck & Panza	2009	Agricultural sciences	Research articles	826416	67.53	9.06	76.59
Konstantakis	2007	Business	Textbooks	600000	85.72	4.66	90.38
Li & Qian	2010	Finance	Reports & speeches	6279702	72.63	10.46	83.09
Cobb & Horst	2004	Anatomy (Learned section Brown corpus)	Research articles	2024	74.85	6.72	81.57
Fraser	2007	Pharmacology	Research articles	185000	60.97	9.47	70.44
Fraser	2007	Pharmacology	Textbook	58413	62.19	6.58	68.77
Fraser	2009	Pharmacology	Research articles	360000	61	9.50	70.50

Despite the differences in size, subject area, and coverage results shown by the specialised corpora summarised in Table 2.2, the lexical coverage of the GSL and AWL in those studies has ranged between 60.97% of tokens (Fraser, 2009) and 85.72% of tokens (Konstantakis, 2007) for the GSL, and 4.66% of tokens (Konstantakis, 2007) and 11.6% of tokens (Cobb & Horst, 2004) for the AWL, with the overall coverage of both lists (i.e., GSL + AWL) between 68.77% of tokens (Fraser, 2007) over the pharmacology textbook corpus and 90.38% of tokens (Konstantakis, 2007) over the business corpus.

Three additional studies on the specialised lexis of three different academic fields: Medicine (Chen & Ge, 2007), Engineering (Ward, 2009) and Applied Linguistics (Vongpumivitch et al., 2009), reported only on the lexical coverage of the AWL (See Table 2.3).

Table 2.3 Lexical coverage of AWL on three specialised corpora

Researchers	Year	Corpus	Text type	Number of tokens	AWL %
Chen & Ge	2007	Medicine	Research articles	190425	10.07
Ward	2009	Engineering	Textbooks	271000	11.30
Vongpumivitch, Huang & Chang	2009	Applied Linguistics	Research articles	1554032	11.17

Even though the three studies in Table 2.3 come from different subject areas and are based on specialised corpora of various sizes, the lexical coverage of the AWL in the three studies differs by little over 1% between the highest coverage (11.3%) and the lowest (10.07%). In Chen & Ge's (2007) study, they identified 471 word families in the AWL with medical meanings and reported a lexical coverage of 10.07%. Chen and Gen (2007) carried out a lexical analysis of frequency and coverage of Coxhead's (2000) AWL over a medical corpus. With this purpose in mind, Chen and Ge (2007) compiled a medical written corpus of 50 research articles comprising a total of 190,425 tokens. The lexical coverage results of their investigation revealed that (1) the AWL covers 10.07% of medical research articles, and (2) out of the 570 AWL word families, 292 of these word families (i.e., 51.2%) are frequently used in medical English. These lexical profile results are similar to other lexical coverage studies (Hyland & Tse, 2007; Li & Qian, 2010) shown in Table 2.2. Moreover, the results of Chen & Ge's (2007) study confirm the importance of academic words in medical English.

In spite of the different sizes of the corpora summarised in Table 2.3, the lexical coverage of the AWL on such specialised corpora is very similar, especially between the Engineering and the Applied Linguistics corpus. In general, the wide variation observed in coverage figures in Table 2.2 and Table 2.3, even among corpora studies related to the medical and health sciences (Fraser, 2007, 2009) has led us to explore further in Chapter 4 the lexical coverage of the GSL, and the AWL over a medical corpus of 5,431,740 tokens.

2.2.3.3 Specialised word lists beyond the GSL and AWL

This section looks at the lexical coverage results of some subject-specific lists that have been designed to be used once learners have mastered the vocabulary of well-known existing word lists like the GSL and the AWL. In an effort to achieve a good lexical

coverage (one that ranges between 95% and 98%) and meet the needs of ESP students, two types of subject-specific word lists have been created: (1) specialised word lists that go beyond other existing lists like the GSL and AWL, and (2) stand-alone specialised word lists. Let us now look at the text coverage results of these new subject-specific word lists developed beyond the GSL, and the AWL. (See Coxhead, 2011 for a summary of the lexical coverage results of the AWL over various specialised corpora).

Following a similar methodology to the one used by Coxhead (2000) to create the AWL, various subject-specific word lists have been developed: a science word list (Coxhead & Hirsh, 2007), medical academic word lists (Hsu, 2013; Wang, Liang, & Ge, 2008), a pharmacology word list (Fraser, 2007), engineering word lists (Mudraya, 2006; Ward, 1999, 2009), business word lists (Hsu, 2014; Konstantakis, 2007), and an agricultural word list (Martínez et al., 2009).

The science, business, pharmacology word lists designed to be used after the GSL and AWL have been mastered include: Coxhead and Hirsh's (2007) Pilot Science List (PSL), Fraser's (2007) pharmacology word list (PWL), and Konstantakis' (2007) business word list (BWL). Table 2.4 below summarises the lexical coverage results of the GSL, AWL and the specialised word list created for each of these three subject areas (i.e., Science, Business, and Pharmacology).

Table 2.4 Lexical coverage of subject-specific word lists created to supplement the GSL and the AWL

Researchers	Year	Corpus type	Number of tokens	Total GSL& AWL %	Specialised list %	Total %
Coxhead & Hirsh	2007	Science research articles & chapters	1761380	80.48	3.79 (PSL)	84.27
Konstantakis	2007	Business textbooks	600000	90.38	2.79 (BWL)	93.17
Fraser	2007	Pharmacology research articles	185000	70.44	12.91 (PWL)	83.35
Fraser	2007	Pharmacology textbook	58413	68.77	14.76 (PWL)	83.53

Coxhead and Hirsh (2007) developed a science list (from now on referred to as the Pilot Science List) for EAP based on a written science corpus of English comprising a total of 2,637,226 tokens. The Pilot Science List research by Coxhead and Hirsh (2007) aims at creating a word list that could help increase the low coverage of the AWL over science

texts (Coxhead, 2000). The complete pilot science corpus used to make the Pilot Science List was developed by using the science subcorpus of Coxhead's written academic English corpus (cited in Coxhead, 2000), and by adding a new science corpus developed by Coxhead and Hirsh (2007). As explained by Coxhead (2000), the science subcorpus used for the AWL includes 875,846 tokens from texts from seven subject areas (i.e., Biology, Chemistry, Computer Science, Geography, Geology, Mathematics, and Physics). Additionally, 1,761,380 tokens from seven other science subject areas (i.e., Agricultural Science, Ecology, Engineering and Technology, Nursing and Midwifery, Sport and Health Sciences, and Veterinary and Animal Sciences) were included in Coxhead and Hirsh's (2007) science corpus. By increasing the size of the corpus, Coxhead and Hirsh (2007) made it more representative of science texts. The following three criteria were considered for selecting the words to be added to the Pilot Science List: (1) *range*[5], i.e., including words that occur at least in seven of the science subject areas, (2) *frequency*, i.e., including words with a frequency of at least 50 or higher in the science corpus, (3) *dispersion*[6], i.e., including words with a dispersion factor of at least 35 (Coxhead & Hirsh, 2007, p. 71). Examples of words in the Pilot Science List are *cell*, *insulin*, *proton*, *vector*, and *zinc*. The coverage of the science corpus is 71.52%, and 8.96% by the GSL and AWL, respectively. As Coxhead and Hirsh (2007, p. 72) reported, the 318 word families in the Pilot Science List cover 3.79% over the science corpus compiled to create this list. Furthermore, the Pilot 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 Pilot Science List. Coxhead and Hirsh's study (2007) 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) Pilot Science List will be used in Chapter 4 of the present thesis when adopting a traditional corpus-based approach

[5] *Range* refers to the occurrences of a word in different sections, files, texts or subject areas of a larger corpus (Coxhead, 2000).
[6] *Dispersion* is a measure of the spread of the distribution of word occurrences across the subfields of a corpus (Coxhead & Hirsh, 2007).

to investigate the vocabulary load of medical texts written in English. (For further evaluation of the Pilot Science list see Chapter 4).

In the field of Business English, Konstantakis (2007) created a Business Word List (BWL) using the Published Material Corpus (PMC) developed by Nelson (2000). As reported by Konstantakis (2007, p. 85), the Published Material Corpus contains around 600,000 tokens from 33 Business course books. The two word selection criteria used by Konstantakis (2007) to choose the 560 word families of the Business Word List were as follows: (1) *range* (including words that occur in at least five course books), (2) *frequency* (including words that occur at least ten times in the whole corpus). Additionally, Konstantakis' (2007) Business Word List only contains words beyond the GSL and AWL. Examples of words in the Business word list are *alliance*, *holdings*, *monetary*, *patent*, and *taxation*. In relation to the exclusion criteria for developing the Business Word List, a group of words (e.g., proper names, numbers, Latin words, nationalities, abbreviations, acronyms and interjections) which Konstantakis (2007, p. 87) considered known or with light learning burden were excluded from this list. The Published Material Corpus covered 85.72% over the GSL, 4.66% over the AWL, and 2.79% over the Business Word List. The cumulative coverage of these three lists (i.e., 85.72%, 4.66%, plus 2.79%) resulted in 93.17%. In an attempt to increase this 93.17% provided by the GSL, AWL and the Business Word List, Konstantakis (2007) developed two new lists: (1) a list of common nationalities which yielded a 0.55% coverage, and (2) a list of abbreviations and acronyms which accounted for 0.30% over the Published Material corpus. In addition, Konstantakis (2007) ran the Published Material Corpus over Nation's (2005) BNC proper name list which provided a 1.63% coverage. By adding the overall coverage of these three lists (i.e., 0.55% by the list of common nationalities, 0.30% by the list of abbreviations and acronyms, and 1.63% by the BNC list of proper names) Konstantakis' (2007) study achieved a good coverage (95.65%) of business course books.

In the field of medical studies, Fraser (2007) looked at the lexical profile of pharmacology texts. In order to achieve this goal, Fraser (2007) compiled a Pharmacology corpus comprising 185,000 tokens from 51 research articles selected from a wide range of pharmacology journals with topics from six pharmacology areas (i.e., cardiovascular pharmacology, autonomic pharmacology, biomedical pharmacology, clinical pharmacology, alimentary pharmacology, and toxicology). Using the Pharmacology corpus

of research articles, Fraser (2007) developed a Pharmacology word list based on the following two criteria: (1) range (i.e., including words that occur in at least 6 articles), and (2) frequency (i.e., including words that occur at least 10 times in at least two articles). Additionally, the 601 word families included in the Pharmacology Word List using these two word selection criteria were words outside the GSL and AWL. In relation to the exclusion criteria to create the Pharmacology Word List, Fraser (2007) followed similar criteria to the ones adopted by Konstantakis (2007). That is, the Pharmacology Word List did not include words that, according to Fraser (2007), had a light learning burden (e.g., proper names, nationalities, numbers, abbreviations and acronyms). The Pharmacology Word List produced the following coverage results: 60.97% by the GSL, 4.46% by the AWL, and 12.91% by the Pharmacology Word List. The 12.61% coverage of the Pharmacology Word List over the Pharmacology corpus shows the good coverage provided by this subject-specific word list. However, a total cumulative coverage of 83.35% by the GSL, AWL, and Pharmacology Word List over the pharmacology texts is still far from an adequate lexical threshold of 95%. As an attempt to increase this 83.35% coverage result, Fraser (2007) developed an additional list of abbreviations and acronyms. Examples of words in the Pharmacology Word List are *abnormality, clone, mutation, neuron, plasma, saline, vitro,* and *toxicity.* This Pharmacology abbreviation list included 140 abbreviations and provided a 4.3% coverage over the Pharmacology corpus. Nevertheless, as Fraser (2007, p. 132) argues, with only a total coverage of 87.66% after adding the 4.31% coverage of the Pharmacology abbreviation list, "we are still, though, a long way from 95% coverage." In order to validate the Pharmacology Word List, Fraser (2007) compiled another Pharmacology corpus consisting of 58,413 tokens from a Pharmacology course book. The coverage results after running this second Pharmacology corpus over the GSL (62.19%), AWL (4.66%), Pharmacology Word List (14.76%) and Pharmacology list of abbreviations (0.58%) confirm the usefulness of the Pharmacology Word List for students of pharmacology who need to become familiar with the lexis of both pharmacology journal articles and textbooks in English.

When we look in Table 2.4 at the overall lexical coverage reached after adding the coverage of the specialised word lists developed beyond the GSL and the AWL, a 93.17% coverage by Konstantakis (2007) has been the closest to an adequate lexical coverage (i.e., 95%). As explained above, a 95.65% was only achieved by Konstantakis (2007) after adding the

coverage results of three more lists, i.e., a word list of business acronyms and abbreviations (with a 0.30% coverage), a list of the most common nationalities (with a 0.55% coverage), and Nation's (2005) BNC list of proper nouns (with a 1.63% coverage). No study has achieved so far a good lexical coverage (of at least 95%) solely on medical texts beyond the GSL and AWL. For this reason, the present study examines the vocabulary load of medical texts written in English beyond existing lists such as the GSL, AWL, and Pilot Science List. The lexical coverage results of the studies summarised in Table 2.4 show the demanding nature of science texts and medical texts especially which require many more words beyond the GSL and AWL to reach a respectable coverage.

2.2.3.4 Specialised word lists beyond the BNC word lists

More recent studies by Hsu (2013, 2014) have calculated the lexical coverage of medical and engineering textbooks and have specifically focused on the development of subject-specific words lists beyond the BNC word family lists (Hsu, 2013) and the BNC/COCA word family lists (Hsu, 2014). Hsu (2013) compiled a large (15,000,000 tokens) corpus of medical textbooks with a twofold purpose in mind: (1) to investigate the number of words beyond the 14,000 BNC word lists needed to achieve an optimal lexical threshold (98%), and (2) to estimate the coverage of a Medical Word List (MWL) beyond the first 3,000 BNC word family lists. For reaching a 98% coverage, Hsu (2013) made two medical word lists (i.e., a word list of technical medical terms with 3,474 entries, and a list of 1,427 medical abbreviations). These two medical lists were created using words outside the 14,000 BNC word family lists plus proper nouns. It was unclear what unit of counting was used for grouping the items in these two medical word lists. In relation to the second purpose of Hsu's (2013) investigation, a Medical Word List (containing 525 word families) from a corpus of medical textbooks was made.

Hsu's (2013) Medical Word List includes similar kinds of words (i.e., general and academic words with high frequency of occurrence in medical texts) to another medical word list, that is, Wang, Liang and Ge's (2008) Medical Academic Word List (MAWL) which contains 623 word families. According to Hsu (2013), one of the main differences between the Medical Word List and Medical Academic Word List lies in the type of medical texts used to build the specialised corpora, namely, medical textbooks for Hsu's (2013) corpus, and medical research articles in the case of Wang, Liang and Ge's (2008) corpus. An

additional difference between these two medical word lists (i.e., Medical Word List and Medical Academic Word List) is that both lists were created beyond different existing word lists. That is, while Hsu's (2013) Medical Word List comprises words outside the first 3,000 BNC lists, Wang, Liang and Ge's (2008) Medical Academic Word List includes words beyond the GSL1 and GSL2 (West, 1953). In relation to the lexical text coverage results, both lists, Medical Word List and Medical Academic Word List, provide similar coverage over medical texts (i.e., 10.72% and 12.24%, respectively). For the present investigation, a different methodology to the one adopted by Hsu (2013), and Wang, Liang and Ge (2008) has been followed to identify various possible lexical shortcuts for English for Medical Purposes vocabulary learning and teaching. This identification was aimed at providing English for Medical Purposes teachers and students with the most frequent and relevant lexis that occurs in medical texts. Additionally, the present thesis estimates the lexical coverage of medical texts using the BNC/COCA word family lists, but develops a different methodology to the one used by Hsu (2013).

Table 2.5 Lexical coverage of first and second 1,000 BNC and BNC/COCA word lists over Hsu's (2013, 2014) medical and engineering corpora

Year	Corpus type	Number of tokens	BNC 1 and 2 %	BNC 1, 2, 3 %
2013	Medicine textbooks	15016553	70.68	81.40
2014	Engineering textbooks	4575413	78.15	88.63

Table 2.5 shows in column four the coverage results of the first and second 1,000 BNC/COCA lists over the medical and engineering corpus. For both studies, this is the point after which new subject-specific lists have been added.

In the particular case of the new medical word lists created by Hsu (2013), in order to be able to achieve a 99.7% lexical coverage over the medical corpus of textbooks, she added to the first two 1,000 BNC list coverage (70.68%), the coverage results of four new medical word lists she compiled, namely, a list of 575 medical academic words (12.42%), a list of 3,474 medical terms (14.39%), a list of 1,427 medical abbreviations (0.77%), and a list of 5,952 proper nouns (1.44%). As indicated in Table 2.5, the inclusion of these four medical word lists to the first two 1,000 BNC lists resulted in a total coverage of 99.07% over Hsu's (2013) medical corpus. (See Table 2.6).

In the engineering corpus, in order to be able to reach a 95% coverage, Hsu (2014) added to the lexical coverage result of the first and second 1,000 BNC/COCA lists (78.15%), the coverage of the BNC/COCA proper nouns list (1.17%), the BNC/COCA transparent compounds list (0.29%), the BNC/COCA abbreviations list (1.09%), and a new Engineering English Word List (EEWL) list (14.3%) that Hsu (2014) developed and grouped into word families.

Table 2.6 Lexical coverage of Hsu's (2013, 2014) subject-specific word lists created to supplement the lexical coverage of the first and second 1,000 BNC lists and BNC/COCA lists, respectively

Researchers	Year	Corpus type	Number of tokens	Number of new lists	Headwords in new lists	Total %
Hsu	2013	Medical textbooks	15016553	4	11448	99.70
Hsu	2014	Engineering textbooks	4575413	1	729	95

A notable feature of Hsu's research is the very large corpus sizes (see column 4 of Table 2.6) that she used. Her medical lists however were not as extensive as the ones used in this thesis.

In the next section, some specialised vocabulary studies that have created stand-alone word lists are mentioned. These discipline-specific word lists have been created with the goal of providing a lexical shortcut for accelerating the learning and growth of vocabulary from particular subject fields.

2.2.3.5 Stand-alone specialised word lists

Attempts have been made by vocabulary researchers to create specialised word lists like Fraser's (2009) Pharmacology Word List (PWL), Ward's (1999, 2009) engineering word lists, i.e., Engineering Word List (EWL; 1999) and Basic Engineering List (BEL; 2009), and Wang, Liang and Ge's (2008) Medical Academic Word List (MAWL). These word lists include only words related to a specific subject area such as Pharmacology, Engineering, and Medicine.

Note in Table 2.7 that most of the studies reported on in this section used rather small corpus sizes of less than one million tokens.

Table 2.7 Lexical coverage results of standalone specialised word lists organised by year

Researchers	Year	Corpus type	Number of tokens	Headwords	Lists	Total coverage of new lists%
Ward	1999	Engineering textbooks	1000000	2000	EWL	95.70
Ward	1999	Mechanics textbooks	16000	3000	EWL	97.80
Wang, Liang & Ge	2008	Medicine research articles	1093011	623	MAWL	12.24
Ward	2009	Engineering textbooks	271000	299	BEL	16.40
Fraser	2009	Pharmacology research articles	360000	2000	PWL	89.10
Fraser	2009	Pharmacology textbooks	60000	2000	PWL	86.30

As summarised in Table 2.7, the specialised word lists created by Ward (1999, 2009) and Fraser (2009) include subject-specific vocabulary that can be used by L2 learners who, in most cases, are still mastering general high-frequency words in English. By focusing only on the words that actually occur in these particular subject areas, the subject-specific word lists in Table 2.7 provide a vocabulary learning shortcut for ESP learners. That is, instead of learning the words of English according to frequency levels, which may include words that are immediately useful for their study and words that are not, the learners go directly into learning subject-related vocabulary. This approach has the advantage of satisfying an immediate need, but has the disadvantage of a restrictively narrow focus that excludes other uses of the language. However, generally the goal of English for Specific Purposes is to provide a narrow focus in the interests of efficiency.

The Medical Academic Word List (MAWL) designed by Wang, Liang and Ge (2008) includes frequent academic words from the medical texts. This is the only stand-alone list in Table 2.7 that excluded the most frequent 2,000 word families in English (i.e., the GSL by West, 1953). Wang, Liang and Ge (2008) conducted a corpus analysis of 288 medical research articles. The lexical profile results provided by Wang, Liang and Ge's (2008) medical corpus of 1,093,011 tokens were used to create a Medical Academic Word List (MAWL) of 623 word families. This list was designed using similar word selection criteria to Coxhead's (2000) AWL, i.e., (1) *specialised occurrence*, (2) *range*, and (3) *frequency*. The MAWL yielded a 12.24% lexical coverage of medical research articles and showed

55% overlap with the AWL word families. From a semantic point of view, the MAWL includes words with general academic meaning (e.g., *previous, similar, data*), and words with more specialised medical meaning (e.g., *vein, lesion, cell*). The MAWL confirms the important role played by the high-frequency academic and specialised words in medical texts.

The unit of counting for most of the lists in Table 2.7 is the word family, with Ward's (2009) Basic Engineering List (BEL) being the only list that counted the lexical items as individual word types. 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 Chung & Nation, 2003 for further discussion on units of counting). Ward (2009) compiled a corpus of engineering textbooks from five engineering subdisciplines (i.e., chemical, civil, electrical, industrial and mechanical engineering) comprising a total of 250,000 tokens. Ward (2009) used this engineering corpus to develop a Basic Engineering List (BEL) for foundation engineering students in Thailand. According to Ward (2009, p. 173), the Basic Engineering List was created "to identify the vocabulary frequent in a wider representation of engineering subdisciplines" and provide engineering students with the specific lexical knowledge required by these engineering textbooks. This list contains the most frequent 299 content words (word types) with a frequency of at least 25 in the complete engineering corpus. The 299 word types of the Basic Engineering List provide a lexical coverage of 17.2%, 15.6% and 21% over three other engineering textbooks, and 11.3% over the engineering corpus used to create this list. These results show a good coverage of the Basic Engineering List over engineering textbooks. However, even for those L2 learners who may know the 299 word types in the Basic Engineering List, these engineering students still need to gain a lexical knowledge of at least 74% or 77% in order to reach an adequate (95%) or optimal (98%) lexical comprehension of engineering textbooks. Examples of words in the Basic Engineering List are *equation, process, show, system, temperature*. These examples indicate that the Basic Engineering List includes a mix of the most frequent 299 general, academic and content specific (technical) words in engineering texts. Generally speaking, Ward's (2009) Basic Engineering List provides a good starting point for engineering students having to read engineering textbooks in English from the beginning of their university degrees.

So far in this literature review, we have looked at L2 vocabulary research that has followed a corpus-based approach to create pedagogical word lists, to calculate the lexical coverage of those word lists, and to estimate the vocabulary load of specialised texts of various specific subject-areas. For vocabulary assessment purposes, corpus analysis can provide a basis for more accurate word lists from which target words can be sampled, taking account of frequency of occurrence, range, and dispersion among other criteria. The corpus-based approach used by vocabulary researchers has offered new insights into the most relevant words in written academic discourse. Corpus analysis yields descriptions of the lexical features of language as it is employed in specific contexts of use, such as academic disciplines. For these reasons, a corpus approach has been chosen in Chapters 4, 5 and 6 of the present investigation to estimate the number of words in medical texts written in English required to reach an appropriate lexical threshold.

In the next section of this review, we look at the role that word lists developed from a corpus-based approach have played in the development of vocabulary size tests and measurement of the vocabulary size of L2 learners. Vocabulary size tests help teachers and course designers see how close learners are to meeting the vocabulary demands of technical texts.

2.3 How do you measure vocabulary size in an L2?

One key component of second language learning is the number of words L2 learners know, their vocabulary size. Vocabulary size tests have been used both by vocabulary researchers and language educators for various purposes, namely, to determine and compare the total number of words known by non-native and native speakers of English (Nation & Coxhead, 2014), to measure learners' vocabulary growth over time (Nation & Beglar, 2007), to design course materials suitable for learners at particular proficiency levels (Nation & Webb, 2011), and to estimate English learners' lexical proficiency for diagnostic purposes, course placement or research, e.g., in L2 reading development programmes (Nguyen & Nation, 2011). Vocabulary size tests differ from other tests of vocabulary knowledge in that they are based on samples drawn from a known population of words so that a learner's score on the test can be converted to an estimate of vocabulary size by multiplying the score by the ratio of sample size to total population size (Nation, 2013).

Vocabulary size tests designed for second language research and instructional purposes are discrete measures of word knowledge that provide an estimate of the total number (also known as 'quantity' or 'breadth') of words known by L2 learners (Read, 2000). A variety of test formats and methods have been used for measuring vocabulary size receptively in L2 settings. A receptive vocabulary test is intended to measure the learners' ability to recognise the spoken or written form of a word and recall or recognise its meaning. That is, such tests measure learners' recognition and knowledge of words (Read, 2000; Schmitt, 2008). Developing tests to measure the vocabulary size of L2 learners has involved sampling from word frequency lists or large dictionaries to obtain the words required to construct the tests.

The number of English words native and non-native university students know has been the subject of continuous research for many decades (Nation, 2013; Nation & Coxhead, 2014). Next, we refer to some of the most influential receptive tests for measuring the written vocabulary size of L2 adult learners of English in several higher education institutions around the world. The discussion is organised on the basis of the test formats used, namely, the yes/no format, the matching format, the multiple-choice format, and the multiple-choice translation format. Pen-and-paper versions and more recent computer-based versions of these test formats are available.

2.3.1 The Yes/No format in vocabulary size testing

Most of the long-established Yes/No tests in L2 research have been developed by Meara and his colleagues (Meara & Buxton, 1987; Meara & Jones, 1988, 1990; Meara & Milton, 2003; Miralpeix & Meara, 2013) over more than two decades. The design of Meara and his associates' Yes/No tests builds initially on previous work on L1 reading research and vocabulary knowledge by Anderson and Freebody (1982).

In general, Yes/No tests present a list of individual words chosen from different frequency levels. The words are presented in isolation, that is, without a syntactic or semantic context. While taking the test, L2 learners are required to either only select the words they know, or click on 'Y' or 'N' depending on whether they recognise the word or not. A percentage (normally ranging between 20% and 50%) of non-words or pseudowords is included to control for guessing, by adjusting the scores to correct for over-reporting of the test takers'

word knowledge (Read, 2007). Non-words are primarily used in Yes/No tests as an attempt to validate test-takers' consistency in reporting correct rejections (i.e., a 'No' response to a non-word). Examples of non-words used in Pellicer-Sánchez and Schmitt's (2012, p. 20) Yes/No test are *berrow, bodelate, cambule, haque, pring*. As Sevigny and Ramonda (2012, p. 702) note, "non-words follow the phonetic rules of English and provide a window into determining whether students are honestly stating their familiarity or unfamiliarity with vocabulary items." The simplicity of the format (i.e., Yes/No tests are context independent, easy to construct and quick to take, easily computerised) makes them attractive and cost effective.

Meara and Jones (1990) developed a computer-based version of the Yes/No test for the Eurocentres language schools in continental Europe and the UK. The Eurocentres Vocabulary Size Test (EVST) was devised to serve as a placement measure. It included a sample of real words to provide "an estimate of a learner's vocabulary size using a graded sample of words covering numerous frequency levels." (Read, 2000, p. 126). About one-third of the words in the test were non-words. The EVST, like other Yes-No tests, is not meant to measure deep lexical knowledge, but simply written receptive knowledge of a word by recognising it (Eyckmans, 2004).

The Yes/No format was also selected for the Vocabulary Size Placement Test (VSPT) devised by Meara and his associates at Swansea University for the European DIALANG project (www.dialang.org) with the support of the European Commission (cited in Alderson, 2005, pp. 79–96). The Vocabulary Size Placement Test has been used as part of the DIALANG computer-based diagnostic language testing system for pre-assessing language ability in 14 European languages. The main purpose of the Vocabulary Size Placement Test is to be a measurement tool that provides a pre-estimate of the test takers' language proficiency and then assigns them to the DIALANG battery of tests that best matches their language proficiency. As Alderson (2005, p. 88) stated, the results of the Vocabulary Size Placement Test have provided enough evidence to claim that "the size of one's vocabulary is relevant to one's performance on any language test", and that the Vocabulary Size Placement Test is "a quick and reliable placement procedure for more detailed diagnoses of different aspects of language ability."

There is still no agreement among L2 vocabulary researchers (Beeckmans, Eyckmans, Janssens, Dufranne, & Velde, 2001; Harrington & Carey, 2009; Mochida & Harrington, 2006; Pellicer-Sánchez & Schmitt, 2012; Shillaw, 1996, 2009; Zhang & Lu, 2014) on the best scoring methodology to follow when adjusting the scores of Yes/No tests based on responses to the non-words in the test. In particular, there is lack of consensus on dealing with the over-reporting of non-words; that is, saying they are known words. A relevant question about Yes/No tests has frequently been: What can be done to prevent over-reporting of word knowledge, and checking 'Yes' (i.e., false alarm) for non-words or words the test takers do not actually know?

Recent computer-based vocabulary studies (Harrington & Carey, 2009; Miralpeix & Meara, 2013; Pellicer-Sánchez & Schmitt, 2012) have looked at the relationship between speed and accuracy of the responses in Yes/No tests. The results of these studies have no clear evidence of a strong correlation between response time and the vocabulary size of the L2 participants taking part in those studies.

In sum, Yes/No tests provide a rough estimate of how many words a test taker can recognise. Recognition of a word, however, does not really guarantee that L2 learners can use the words they claim to know. One of the limitations of Yes/No tests might be the presentation of the test items as a list of words without context, since non-defining contexts can help in orienting learners to the part of speech of words. Perhaps the greatest weakness of the Yes/No tests is their face validity (Nation & Webb, 2011). Learners are assessed on their word knowledge without having to show that they know the meaning of each word. Although the tests work well, this face validity issue means that teachers are often reluctant to use them. This face validity problem is not an issue in tests where learners are required to produce or choose a meaning as in the matching format.

2.3.2 The matching format in vocabulary size testing

A matching format was adopted in the design (Nation, 1983, 1990) and revised versions (Schmitt, Schmitt, & Clapham, 2001) of the Vocabulary Levels Test. In 1993 Norbert Schmitt revised the first version of the Vocabulary Levels Test and developed three more versions of the test. A few years later, Schmitt, Schmitt and Clapham (2001, p. 79) conducted a validation study of the revised versions of the Vocabulary Levels Test and

found that “they can provide valid results and produce similar, if not truly equivalent, scores.”

The Vocabulary Levels Test required test takers to match the tested words with their synonyms or definitions. This test provided general information about the specific five frequency levels assessed in the test. The words included in the original Vocabulary Levels Test come from five word frequency levels – the first 2,000, 3,000, 5,000, the University word list, and 10,000 level. Here is an item from the first 2,000 word frequency level of the Vocabulary Levels Test illustrating the use of six words and three meanings in each section (Nation, 1990, p. 266):

1. accident
2. choice _____ having a high opinion of yourself
3. debt _____ something you must pay
4. fortune _____ loud, deep sound
5. pride
6. roar

The Vocabulary Levels Test was devised as a diagnostic vocabulary measure for estimating L2 learners’ receptive word knowledge of each of the frequency levels assessed in the test. Although, the Vocabulary Levels Test was not originally designed by Nation (1983) as a measure of vocabulary size, it has been often used to determine the number of words known (i.e. vocabulary size) and learnt (i.e., vocabulary growth). It is not rare to find vocabulary studies (e.g., Cobb & Horst, 2004; Haliza, Ibrahim, Othman, Sarudin, & Muhamad, 2013; L. Li & MacGregor, 2010; Nacera, 2010; Nemati, 2010; Zhang & Lu, 2014) in various L2 educational settings around the world that have used the Vocabulary Levels Test as a size test. The Vocabulary Levels Test is however more likely to be less sensitive as a measure of vocabulary size than, for example, Nation and Beglar’s (2007) Vocabulary Size Test. In relation to the purpose of the Vocabulary Levels Test, Paul Nation (2013, p. 36), its creator, has asserted that the test is “a diagnostic test” that “does not measure how many words someone knows.”

2.3.3 The multiple-choice format in vocabulary size testing

The multiple-choice format has been one of the most widely used test formats in standardised vocabulary testing. It was also the test format chosen by Paul Nation and David Beglar in 2007 to design a vocabulary test that could serve as a size measure of non-native speakers' word knowledge.

The multiple-choice format of the Vocabulary Size Test (VST) created by Nation and Beglar (2007) consists of 140 items (i.e., 10 items randomly drawn from each of the fourteen 1,000 word levels). More recently, two parallel versions of the VST, each with a total of 100 items (i.e., 5 items randomly drawn from each of the twenty 1,000 word levels) were developed (Nation & Coxhead, 2014). Each multiple choice item includes a tested word (target word) in a simple non-defining context, and four options (the correct answer and three distractors). The VST is based on word frequency lists (Nation, 2006) developed from frequency data provided by the British National Corpus (BNC) and the lists range from the first 1,000 to the fourteenth 1,000 word families for the 140-item version of the VST (Nation & Beglar, 2007), and from the first 1,000 to the twentieth 1,000 word families for the two more recent 100-item versions of this test (Nation & Coxhead, 2014). Here is a sample item from the fourteenth 1,000 level of Nation and Beglar's (2007) VST.

	canonical: These are canonical examples.	
	a.	examples which break the usual rules
	b.	examples taken from a religious book
√	**c.**	**regular and widely accepted examples**
	d.	examples discovered very recently

Nation and Beglar (2007) selected the multiple-choice format over other test formats commonly used in vocabulary size testing, because it allows the presentation of the target words and the four options in a non-defining context. According to Read (2000), contextualised words provide a much richer environment and may enhance the learner's awareness of the usage of the words.

Even though the receptive vocabulary tests above mentioned (i.e, the EVST, Vocabulary Size Placement Test, Vocabulary Levels Test) have proven to be valid, reliable, and practical receptive measures of word knowledge (Alderson, 2005; Read, 1988; Schmitt et al., 2001), the VST was chosen as the vocabulary size measure for the present study (see

Chapter 3). One of the main reasons for selecting the VST is because it was designed specifically to measure vocabulary size. In addition, the multiple-choice format provides more than a word recognition count by presenting the target words in a non-defining syntactic context. This linguistic context also indicates how the target word functions (i.e., its word class, and grammatical patterns) allowing learners to verify if they know the tested words.

2.3.3.1 Bilingual multiple-choice format of the VST

As explained by Nation and Coxhead (2014), the 14,000 word version of the VST has been translated into several languages, including Japanese, Korean, Persian, Russian, Spanish, Thai, Vietnamese, simplified Chinese (Putonghua in mainland China), and traditional Chinese (Mandarin in Hong Kong and Taiwan). Here is a sample item from the fourteenth 1,000 level of the Spanish version.

	canonical: These are canonical examples.	
	a.	ejemplos que no siguen las reglas convencionales
	b.	ejemplos tomados de un libro religioso
√	**c.**	**ejemplos comunes y ampliamente aceptados**
	d.	ejemplos descubiertos muy recientemente

Currently five of these bilingual versions (Japanese, Korean, Russian, Vietnamese, and simplified Chinese) are available for free download from Paul Nation's website (http://www.victoria.ac.nz/lals/about/staff/paul-nation). Likewise, seven of these bilingual versions (i.e., Japanese, Korean, Russian, Spanish, Vietnamese, simplified Chinese, and traditional Chinese) can be taken online from http://my.vocabularysize.com/

A year after a validation study of the VST was conducted (Beglar, 2010), the results of the administration of bilingual versions of the VST in Vietnam (Nguyen & Nation, 2011), Iran (Karami, 2012), and Russia (Elgort, 2013) were published. These three studies consistently reported a mean vocabulary size of around 6,000 word families. Elgort (2013) was the first author to publish results on a simultaneous administration of the bilingual and monolingual VST to a group of L2 English learners.

The bilingual VSTs employ both the target language (English in the stimulus sentence presenting the target word in a non-defining context), and the test taker's native language

(in the four options or alternative items). There is only one correct L1 equivalent for each target word. The test taker is simply required to identify the L1 equivalent (synonym or short phrase with a definition or description) that best corresponds to the target word. The other three options are distractors which are either related to the context of the sentence or to the form of the target word. The choices all belong to the same word class. This is done to avoid giving any clues regarding the meaning of the word based on part of speech. (See Chapter 3 for a discussion of the translation principles of bilingual versions of the VST). Here is a sample item from the second 1,000 level of the bilingual Russian (Elgort, 2013) and Vietnamese (Nguyen & Nation, 2011) versions of the VST.

	Monolingual VST item		**Russian VST item**		**Vietnamese VST item**	
	stone: He sat on a stone.		**stone**: He sat on a stone.		**stone**: He sat on a stone.	
√	**a.**	**hard thing**	**a.**	**камень**	**a.**	**hòn đá**
	b.	kind of chair	b.	стул	b.	cái ghế
	c.	soft thing on the floor	c.	ковёр	c.	tấm thảm
	d.	part of a tree	d.	ветка	d.	cành cây

The presentation of the target words in context is consistent with Chapelle's (1994, p. 164) interactionalist definition of the relativity of vocabulary meaning and the need to assess it in relation to the context in which it is used, rather than in absolute terms. Likewise, giving the target words in a sentence context promotes comprehension and helps to disambiguate the meaning of polysemous words.

The various test formats described above include the vocabulary size measures commonly used in second language studies. Nowadays, tests like the Vocabulary Size Placement Test and the VST are well-established measures of vocabulary size for L2 learners of English. In general, the discrete measures of word knowledge mentioned in this section have proven to be useful word measurement tools for vocabulary researchers and L2 teachers alike.

Following similar randomisation and test assignment procedures to the ones used by Elgort (2013) for the administration of both the monolingual and bilingual VST in Russia, this thesis presents the results of the administration of the monolingual and bilingual Spanish VSTs to L1 Spanish university students about to start their first ESP courses in higher education institutions in a Spanish speaking country (Venezuela). The purpose of the administration of the VST to this group of participants is to investigate whether the bilingual Spanish VST could be, as in the case of L1 Russian speakers, a better indicator of

the vocabulary size of L1 Spanish speakers than the monolingual Vocabulary Size Test, and to find the vocabulary size of medical students so that the vocabulary load of their medical texts can be more usefully assessed.

2.4 What levels of vocabulary have been considered when estimating the vocabulary load of academic texts?

Since the main purpose of this investigation is to estimate the number of words (vocabulary load) that learners of English for Medical Purposes need to know in order to be able to meet the lexical demands of medical texts written in English, this section of the literature review refers to two main classifications (Nation, 2013; Schmitt & Schmitt, 2012) currently used to identify the different levels of vocabulary of academic texts. Frequency and text type (i.e., general, academic, scientific, technical or specialised) are the two main criteria used to classify words, and establish the two classifications on the levels of vocabulary discussed below.

2.4.1 Schmitt and Schmitt's (2012) levels of vocabulary

Schmitt and Schmitt (2012) proposed a frequency-based classification consisting of the following three bands or levels, namely, high-frequency, mid-frequency, and low-frequency words. High-frequency words include the top 3,000 most frequent word families. Mid-frequency words comprise those word families which exist between the 4,000 and the 9,000 frequency bands. Low-frequency words refer to those word families beyond the 9,000 frequency level. The introduction of the mid-frequency tier in Schmitt and Schmitt's (2012) classification has served to stress the importance of mid-frequency vocabulary and of words beyond the 3,000 most frequent words of the language. High-frequency vocabulary and mid-frequency vocabulary are seen as being major learning goals for foreign language learners, because these 9,000 word families give them enough vocabulary to read non-specialised unsimplified texts such as newspapers, novels and magazines (Coxhead, 2012; Nation, 2006).

Vocabulary researchers (Douglas, 2014; Minshall, 2013; Nation, 2013; Nation & Anthony, 2013) are already looking at the role played by Schmitt and Schmitt's (2012) addition of the mid-frequency tier to the classification of vocabulary by frequency bands. Extending

the classification, from only having high and low-frequency words to having three frequency levels (i.e., high, mid and low), has also raised awareness of the instructional implications of designing language course materials that provide L2 vocabulary learners with opportunities to meet, practice, and learn words that are essential for reaching an appropriate lexical threshold.

This frequency band classification proposed by Schmitt and Schmitt (2012) is also used in Chapters 5 and 6 of this thesis as one of the methodological frameworks chosen to estimate the vocabulary load of medical texts written in English.

2.4.2 Nation's (2013) levels of vocabulary

Nation's (2013) classification of the levels of vocabulary, which is the most comprehensive classification of its kind up to the present, is both a frequency and text type based classification that includes the following frequency levels as in Schmitt and Schmitt (i.e., high, mid, and low-frequency vocabulary), and text types (i.e., academic, and technical vocabulary). The kinds of vocabulary according to Nation's classification (2001, 2013) – initially presented in 2001 and then revised in 2013 – are described below.

2.4.2.1 High-frequency vocabulary

High-frequency vocabulary comprises a small number of words (two to three thousand word families) with a high frequency and a wide range of occurrence in all kinds of texts, registers and uses of the language. As Nation (2013, p. 22) states, a great amount of written and spoken language is formed by a relatively small number of words (i.e., the most frequent 2,000 word families in English). These words occur so often in the language in every kind of text that they are useful to all learners. For this reason, high-frequency words are very common lexical items that language learners in general and ESP learners in particular should focus on. Here are some high-frequency function words (e.g., *about, by, during, the*) and content words (e.g., *begin, fire, garden, hope*) in the first 2,000 BNC/COCA (Nation, 2012) word family lists.

The high-frequency words are characterised by including most function words (around 170 word families) and many content words of the language, being necessary for all language

learners, being readily accessible to the lay reader (Ward, 1999), covering a large proportion of the running words of the language, i.e., having a wide lexical text coverage, and making up about 80% (Nation, 2013, p. 18) of the running words in written academic texts.

2.4.2.2 Mid-frequency vocabulary

Nation (2013) added a mid-frequency tier (i.e., word families between the 4,000 and the 9,000 frequency levels) to his classification of the levels of vocabulary and also to one of his most recent studies on graded readers (Nation & Anthony, 2013). This classification corresponds to the frequency tier initially proposed by Schmitt and Schmitt (2012). Mid-frequency words are characterised by including 6,000 word families which range between the 4,000 and the 9,000 word frequency levels, consisting of mainly general purpose words with medium frequency, including some useful word families that nearly made it into the high-frequency word lists, and occurring with lower frequency than the high-frequency words, but with higher frequency and wider range than the low-frequency words. Examples of mid-frequency words in the BNC/COCA (Nation, 2012) word family lists are *cactus, commence, glacial, habitual, pear*.

2.4.2.3 Low-frequency vocabulary

Low-frequency vocabulary consists of a large number of words (many thousands of words) that generally have a low frequency and an unpredictable range of occurrence across texts, registers and uses of the language. These are words infrequently met and used, but may include high-frequency specialist vocabulary. According to Nation (2013), low-frequency words are characterised by consisting of tens of thousands of words beyond the 9,000 frequency range, including words that are not high-frequency or mid-frequency words. Examples of low-frequency words in the BNC/COCA (Nation, 2012) word family lists are *avium, benzol, bejeweled, dynast, octal, whereon.*

Because low-frequency words typically cover a small proportion of any text, they have not been given very much attention by vocabulary researchers and ESP teachers. Since this study seeks to provide an integrative view of the levels of vocabulary that form academic texts, low-frequency words are obviously a kind of vocabulary that will also need to be

considered. The above description of the levels of vocabulary that constitute academic texts points to the need for an approach that considers the different kinds of vocabulary that constitute academic texts at various frequency levels.

2.4.2.4 Academic vocabulary

Academic vocabulary (also known as sub-technical vocabulary by Baker, 1988) refers to a small number of words with higher frequency and a wider range of occurrence in a wide variety of academic texts, but which are not so common in non-academic texts. According to Coxhead (2000), academic words are the words all learners require, regardless of their chosen course of study at tertiary level. That is, these kinds of words are essential for tertiary level learners.

Academic words are characterised by

- Having a lower frequency than high-frequency words.
- Having a higher frequency and a wider range in academic texts than in other kinds of texts.
- Being "supportive of but not central to the topics of the texts in which they occur." (Coxhead, 2000, p. 214).
- Being in the high and mid-frequency levels of the BNC/COCA lists.
- Being outside the most frequent 2,000 words of English (GSL) (Coxhead, 2000).
- Being necessary for language learners at tertiary level (Coxhead, 2000).
- Being an extension of high-frequency words for special purposes.
- Being formal vocabulary.
- Making up around 10% of the running words in academic texts (Coxhead, 2000).

2.4.2.5 Technical vocabulary

Technical vocabulary includes subject-specific words that are related to a particular topic or content area, and are likely to occur only or more frequently in a given specialist area. In this regard, Chung and Nation (2004, p. 259) state that technical terms (also known as technical words) "are likely to occur only in a specialised field or to occur with a much higher frequency in a specialised field than in a different field or in a variety of other texts."

Technical words are characterised by being very common in a specific topic area and context but not so common elsewhere, having 'subject specificity' (Chung & Nation, 2004, p. 255) that is, having a meaning that is closely related to a particular topic or subject area, consisting of single words and multiword units (Chung, 2003; Ward, 1999) occurring within a limited area (narrow band) of the text, that is, being clustered (Ward, 1999), including many words of Latin and Greek origin, and making up from around 20% to around 30% (Chung & Nation, 2003) of the running words (tokens) in academic texts.

Existing pedagogical vocabulary lists of general high-frequency words (West's GSL) and academic words (Coxhead's AWL), although widely used at present, cannot provide a complete coverage of the kinds of vocabulary in academic texts. This happens particularly because the GSL and the AWL were not designed to identify all the different kinds of vocabulary of academic texts. For this reason, a more inclusive approach to identify the various levels of vocabulary that occur in academic texts could provide a clearer picture of the vocabulary demands of academic texts.

The present research looks at the lexical profile of medical texts and explores in detail 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 three topic-based word lists (AWL, Pilot Science List, 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 studied in this thesis.

2.5 How can we identify technical vocabulary?

A number of corpus studies (Chung & Nation, 2003; Coxhead & Hirsh, 2007; Coxhead, 2000; Durrant, 2009; Fraser, 2007, 2009; Gardner & Davies, 2014; Hsu, 2013, 2014; Konstantakis, 2007, 2010; Wang et al., 2008; Ward, 1999, 2009) have shown that using corpus data analysis is a suitable way to identify the levels of vocabulary, and the technical words of academic and specialised texts.

Recently, Gardner and Davies (2014) proposed four different statistically based criteria – ratio, range, dispersion, and discipline measures – for distinguishing technical words from core academic words and general high-frequency words. They used an academic corpus of

120 million tokens of written English, i.e., a subsection of the 425-million-token Corpus of Contemporary American English (COCA; Davies, 2014). Word proportion and frequency in academic versus non-academic texts are the two key concepts considered to apply the four word selection criteria above mentioned. The AVL (Gardner & Davies, 2014) when compared with the AWL (Coxhead, 2000) represents different ways of identifying academic vocabulary. While the AWL was created beyond two existing high-frequency general word lists (i.e., the GSL1 and GSL2; West 1953) using criteria of specialised occurrence, range and frequency. In the case of the AVL, it was not built on top on any existing general word list. As previously mentioned, the AVL was actually created taking into account the four criteria of ratio (as a way of distinguishing general high-frequency vocabulary from core-academic vocabulary), range, dispersion and discipline measure (as a way of identifying subject-specific words). Gardner and Davies' (2014) approach for differentiating general, academic, and technical words resulted in the AVL having twice the coverage (14%) of the AWL (7%) over the academic sections of two academic corpora (i.e. the Corpus of Contemporary American English, and the British National Corpus). Additionally, the higher coverage of the AVL can be attributed to the high percentage (over 40%) of general high-frequency word families from the GSL or BNC/COCA lists the AVL list includes.

Chung and Nation (2003, 2004) have carried out the most comprehensive investigation of ways of identifying technical vocabulary in specialised texts. They tried four different approaches – (1) a semantic rating scale, (2) a technical dictionary, (3) typographical clues, and (4) a corpus comparison of frequencies, and range of occurrence – to identify technical words in an anatomy text and in an applied linguistics text. As part of their investigation, they concluded that (1) the rating scale approach was the most valid and time-consuming, (2) the technical dictionary approach provided the most direct way of identifying technical words but was subject to the availability of subject-specific dictionaries, and the classification criteria used to select the technical words included, (3) the clue-based approach was not practical because it required a lot of decision making and extra judgement and showed the lowest accuracy rate, and (4) the corpus comparison approach was the most appropriate in terms of practicality for identifying technical words. Chung and Nation's (2003, 2004) studies highlighted the importance of technical vocabulary and pointed to the need for re-examining the role played by technical vocabulary in specialised texts. For the

present study, the corpus comparison approach and an adaptation of Chung and Nation's (2003, 2004) rating scale approach are the two approaches followed for identifying technical words in medical written texts.

As Table 2.8 shows, Chung and Nation's (2003) semantic rating scale classifies words in specialised texts as technical when they meet the criteria summarised under steps 3 and 4 of their classification.

Table 2.8 Chung and Nation's (2003) rating scale

Step 1 Words that are not related to a specific specialised topic, field or subject area. Examples in the applied linguistics and anatomy texts are: *after*, *extremely*, *common*, *is*, and *it*.
Step 2 Words minimally related to a given specialised topic or subject field. Examples are: *classrooms*, *co-constructed*, and *measurable* (in the applied linguistics text), *superior*, *structures*, and *protects* (in the anatomy text).
Step 3 Words closely related to a particular discipline or subject of specialisation but also used in general language and applicable to other subject fields. Examples are*: clauses*, *negotiation*, and *glossing* (in the applied linguistics text), *clinical*, *disease*, and *pulse* (in the anatomy text).
Step 4 Words that have a meaning specific to a subject field, have clear restrictions of usage and are not used in general language. Examples are: *hyponomy*, *morphosyntax*, and *phonology* (in the applied linguistics text), *hematopoietic*, *mammary*, and *pedicle*.

Using the rating scale involves looking at each word and deciding where it should be placed on the rating scale. Because a high degree of subjective judgement is involved, such rating may need to be checked by an independent rater. The general question that the rating scale is answering is *How closely related to the subject area is this particular word?*

The use of the above mentioned rating scale by Chung and Nation (2003) to identify technical words in anatomy and applied linguistics revealed that one running word in every three in the anatomy text (31.2%) was a technical word, and one running word in every five in the applied linguistics text (20.6%) was a technical word. These two different results, especially in relation to the percentage of technical words in two specialised fields (i.e., anatomy and applied linguistics), point to the need for re-thinking and re-classifying words that were considered by previous researchers to be high-frequency words (West, 1953), academic words (Coxhead, 2000), general scientific words (Coxhead & Hirsh, 2007) but which according to Chung and Nation's (2003) rating scale are, in fact, technical words in

a particular text. (See Chapter 4 for further evaluation of Chung and Nation's semantic rating scale).

Corpus analyses on various specialised subject areas, such as pharmacology and applied linguistics (Fraser, 2005), engineering (Hsu, 2014), culinary writing (Nordin, Stapa, & Darus, 2012) and food technology (Sonchaiya, Wasuntarasophit, & Chindaprasirt, 2011) have applied Chung and Nation's (2003, 2004) rating scale approach as a way of identifying technical words pertaining to their particular fields of specialisation. In relation to the overall percentage of technical words of these studies, Fraser (2005) and Sonchaiya, Wasuntarasophit and Chindaprasirt (2011) are the two studies that provide comparable estimates of the number of technical words in their particular fields. Their results indicate that there was one technical word in every three running words (35.9%) in the pharmacology text (Fraser, 2005), one technical word in every five running words (15% and 15.24%) in the applied linguistics (Fraser, 2005) and food technology (Sonchaiya et al., 2011) texts, respectively. An adaptation of Chung and Nation's (2003) semantic rating scales is used in the present study.

2.6 What is the technical vocabulary of medical texts?

As can be seen from the information provided in the previous section (2.4) of this literature review, several approaches can be used to identify technical words in medical texts. The use of a dictionary and the use of clues provided by the writer of a text place the decision of determining what is a technical word in the hands of others. This is not very satisfactory, as the criteria used are not described and are thus not replicable. The use of a rating scale comes the closest to touching the nature of technical words, namely, their strong association with the field of study. As Chung (2003) showed, the range and frequency-based corpus comparison approach yields similar results to the rating scale approach and is much more practical.

Vocabulary researchers (e.g., Chen & Ge, 2007; Chung & Nation, 2003, 2004; Fraser, 2005, 2009; Hsu, 2013; Wang et al., 2008) differ on where to draw the line between technical and non-technical vocabulary. This issue relates to their view of technical vocabulary. The popular view of technical vocabulary is that it consists of words that are precise, special and unusual because of their relationship with a technical area (see Nordin

et al., 2012; Ward, 2007). That is, they are words that only a specialist would use. Chung and Nation's (2003) approach included commonly known words as technical words if those words were closely related in meaning to a specialist area. So, words such as *neck, chest, arm, heart, ill* would be classified as technical words in medicine because they were closely related in meaning to medicine even though they were widely known high-frequency words. For some researchers (e.g., Chen & Ge, 2007; Wang et al., 2008) using word lists like the GSL and AWL for their lexical profile analyses, the technical vocabulary of medical texts refers simply to those words beyond these well-known word lists. This problem underlines the importance of distinguishing frequency-based lists (high, mid, low) from topic-based lists.

When a rating scale approach has been followed to make decisions on the number of technical words in medical texts, different degrees and thresholds of technicality have been established. While, at one end of the continuum, vocabulary research on medical texts (Fraser, 2009) may have adopted what could be considered as an inclusive view on where to draw the line between technical and general academic vocabulary. Examples of medical words in Fraser's (2009) Pharmacology Word List are *growth, treatment, cell, protein, inhibition, regulation, protein, serum, synapse,* and *acetylcholine*. Researchers (e.g., Baker, 1988; Chen & Ge, 2007; Cowan, 1974; Hsu, 2013; Wang et al., 2008) at the other end of the spectrum have a more restricted view of what technical vocabulary is and how to identify and classify technical words in medicine. In those studies, technical words are those words that solely occur in specialised medical texts with a very specific medical meaning that is only well-known by specialists in the medical field. Somewhere in the middle of the medical technicalness spectrum there are studies by Chung and Nation (2003, 2004), and Fraser (2005). Examples of technical words in anatomy texts are *chest, ribs, cavity, organs, breathing, trachea, vertebrae, hematopoietic, viscera,* and *pedicle* (Chung & Nation, 2003, 2004) and in pharmacology are *artery, bacteria, cell, fever, malignant, oocyte, polypeptide, stenosis, transmitter*, and *warfarin* (Fraser, 2005).

The more inclusive view of what a technical word is, which is the view adopted in Chapters 4, 5, and 6 of this thesis, considers the technical words of medical texts as words with different degrees of relatedness to a particular topic in the medical field (this corresponds to Chung and Nation's (2003) steps 2, 3 and 4 of their rating scale), regardless of which frequency level list (high, mid, low) the words occur in. This view of a continuum of

specialty formed by different degrees of technicality has recently been expressed in Hsu's (2013, 2014) studies on the lexis of medical and engineering texts and in Fraser, Davies and Tatsukawa's (2015) investigation on the lexis of medical texts.

Although several authors (Baker, 1988; Chung & Nation, 2003, 2004; Fraser, 2007, 2009; Hsu, 2013; Salager-Meyer, 1983, 1985) have investigated the role played by technical words in medical texts, to the best of the researcher's knowledge, no vocabulary researcher up to now has used a broader approach to identify technical words across the high, mid, and low-frequency bands. Nor have vocabulary researchers investigated the occurrence of technical words in some well-known existing general (GSL, BNC/COCA), academic (AWL), and scientific (Pilot Science List) word lists and created new technical (medical) word lists that include both the medical words found in the existing word lists, and the medical words beyond those lists. Such an approach to the identification of technical words in medical texts is the approach adopted in this thesis to quantify the vocabulary load of medical textbooks written in English.

2.7 Research questions

With the ideas highlighted in this review of the relevant literature in mind, this investigation provides answers to the following research questions:

1. What is the vocabulary size of a group of university students (L1 Spanish speakers) at the start of their first ESP reading course in Spanish medium higher education institutions?
2. Is the bilingual version of the VST a better indicator of vocabulary size for L1 Spanish speakers than the monolingual version?
3. Is there a difference in scores on the bilingual and the monolingual versions of the VST as a result of the language presentation order?
4. What is the vocabulary load of medical textbooks if you know the GSL, the AWL, and the Pilot Science List?
5. What is the vocabulary load of medical textbooks if you know the first 3,000/6,000/9,000 BNC/COCA word family lists?
6. What is the lexical text coverage of the medical word lists created in Chapters 4 and 5 on a new independent medical corpus (i.e., the med2 corpus)?

The following chapters of this thesis provide answers to the above questions. The first three questions are answered in Chapter 3 in which we also explain the administration, and translation procedures of the VST. In Chapters 4, 5, and 6, we calculate the vocabulary load of medical textbooks using existing general and academic word lists, and developing new medical words lists. In the final discussion, we conclude by reflecting on the pedagogical implications involved in designing an English for Medical Purposes reading programme that promotes the independent learning of the lexical items needed to meet the lexical demands of medical textbooks written in English and achieve good reading comprehension of such texts.

Chapter 3: Translation and administration of the Vocabulary Size Test (VST)

3.1 Introduction and justification

The main goal of the present chapter is to identify the vocabulary size of a group of foreign language learners who are native speakers of Spanish. In order to meet this goal, this study used two versions of the Vocabulary Size Test (VST), namely, Nation and Beglar's (2007) monolingual version of the test, and a specially created Spanish version of the same test. This was done for three major reasons: (1) to estimate the vocabulary size of first year undergraduate students taking ESP courses, (2) to investigate whether taking the bilingual VST could be a better indicator of vocabulary size, and (3) to determine whether there is an order effect from taking the monolingual or bilingual versions of the VST first or last.

After describing the making of the test and its administration, this chapter answers these three questions.

1. What is the vocabulary size of the participants?
2. Is the bilingual version of the VST a better indicator of vocabulary size for L1 Spanish speakers than the monolingual version?
3. Is there a difference in scores on the bilingual and the monolingual versions of the VST as a result of the language presentation order?

Estimating learners' vocabulary size is important for the present study because the vocabulary load of academic texts is dependent on the number of words that readers of those texts already know, as well as the nature of the words outside their current knowledge. (See Chapter 2 for further evaluation of the VST).

First of all, the methodology and criteria followed to translate and administer the VST are explained. The methodology section also describes the translation process of the VST.

3.2 Methodology

3.2.1 Development of the Spanish bilingual version of the VST

Currently, there are seven bilingual versions (Japanese, Korean, Mandarin, Persian, Russian, Spanish, and Vietnamese) of Nation and Beglar's (2007) VST available. The Spanish version of the VST was translated by the researcher for the purposes of this study to investigate whether an English/Spanish version of the test would provide a better indicator of the vocabulary size of L1 Spanish speakers than the monolingual version originally developed by Nation and Beglar (2007). Previous studies (Karami, 2012; Nguyen & Nation, 2011) have also examined this assumption by administering only bilingual versions (i.e, English/Persian, English/Vietnamese, respectively) of the VST. The increasing use of bilingual versions of the VST has been based on the belief by vocabulary researchers (Nguyen & Nation, 2011) that the translation of the four options (correct answer, and three distractors) reduces the effect of syntactically complex phrases and subordinated clauses present in some of these four options, and increases the possibility of the test takers easily accessing the meaning of known target words. That is, the bilingual VST has been developed partly with a grammatical reason in mind, and also to investigate whether having the test takers' L1 in the four options, and English in the target words (i.e., tested words or prompts), would provide a more accurate account of the test takers' vocabulary size.

Before explaining the principles applied for the translation of the Spanish bilingual version of the VST, we refer to the principles used in the translation of the Vietnamese (Nguyen & Nation, 2011) and Russian (Elgort, 2013) versions of this test. These translation principles can be summarised as follows:

Translation principle 1: Both the Vietnamese and Russian translations use, as much as possible, an L1 corresponding single word or shorter phrase as the L1 translation equivalent of the original single words, phrases, and definitions in the four L1 multiple choice options translated (i.e., the correct answer and the three distractors). See Example 1 below illustrating this principle with the tested words '*pro*'.

Example 1

Second 1,000 level						
	Monolingual VST item		Vietnamese VST item		Russian VST item	
	pro: He's a pro.		**pro:** He's a pro.		**pro:** He's a pro.	
	a.	Someone who is employed to find out important secrets	a.	thám tử	a.	следователь
	b.	a stupid person	b.	gã ngốc	b.	глупец
	c.	Someone who writes for a newspaper	c.	nhà báo	c.	корреспондент
√	d.	Someone who is paid for playing sport etc	d.	người chơi thể thao chuyên nghiệp	d.	профессионал

In relation to the translation principle 1, Nguyen and Nation (2011), and Elgort (2013) acknowledge that it was not always possible to apply this translation principle. The main reason for this being, according to Nguyen and Nation (2011, p. 91), that "there is no corresponding single word in the learners' first language which corresponds to the choice." As a result of this, when it was not possible to find an L1 corresponding single word or shorter phrase for the original phrases and definitions in the monolingual VST, it was necessary to maintain a longer phrase or definition format for the L1 translation equivalent of the original phrases and definitions. The tested words '*augur*' (Example 2) and '*canonical*' (Example 3) illustrate this idea.

Example 2

Fourteenth 1,000 level						
	Monolingual VST item		Vietnamese VST item		Russian VST item	
	augur: It augured well.		**augur:** It augured well.		**augur:** It augured well.	
√	a.	promised good things for the future	a.	hứa hẹn điều tốt đẹp trong tương lai	a.	предсказывать судьбу
	b.	agreed well with what was expected	b.	xảy ra đúng như dự đoán	b.	соответствовать ожидаемому
	c.	had a colour that looked good with something else	c.	có màu sắc rất hợp	c.	подходит к чему-либо по цвету
	d.	rang with a clear, beautiful sound	d.	có âm thanh rõ và hay	d.	звонить чистым красивым звоном

Example 3

Fourteenth 1,000 level						
	Monolingual VST item		Vietnamese VST item		Russian VST item	
	canonical: These are canonical examples.		**canonical:** These are canonical examples.		**canonical:** These are canonical examples.	
	a.	examples which break the usual rules	a.	phạm luật	a.	противоречащий установившимся правилам
	b.	examples taken from a religious book	b.	lấy từ sách tôn giáo	b.	основанный на реальных событиях
√	c.	regular and widely accepted examples	c.	hợp với qui tắc tiêu chuẩn	c.	привычный и принятый большинством
	d.	examples discovered very recently	d.	mới được khám phá	d.	недавно обнаруженный

Translation principle 2: For the correct answer in the monolingual VST, the Vietnamese and Russian bilingual versions of the VST aimed, where possible, for an L1 translation equivalent of the tested word instead of the L1 translation of the original multiple choice option. That is, in the case of the tested word '*time*' (see Example 4), a Vietnamese *(thời gian)* and Russian (время) translation of *'time'* was used instead of an L1 translation equivalent of the correct multiple choice option, *'hours'*. Translation principle 2 was applied independently of whether the correct answer was a single word (as in Example 4), a short phrase or definition (as in Example 5).

Example 4

First 1,000 level						
	Monolingual VST item		Vietnamese VST item		Russian VST item	
	time: They have a lot of time.		**time:** They have a lot of time.		**time:** They have a lot of time.	
	a.	money	a.	tiền	a.	деньги
	b.	food	b.	thức ăn	b.	еда
√	c.	hours	c.	thời gian **(= time)**	c.	время **(= time)**
	d.	friends	d.	bạn bè	d.	друзья

Example 5

First 1,000 level						
	Monolingual VST item		Vietnamese VST item		Russian VST item	
	poor: We are poor		**poor:** We are poor		**poor:** We are poor	
√	a.	have no money	a.	nghèo **(= poor)**	a.	бедный **(= poor)**
	b.	feel happy	b.	hạnh phúc	b.	счастливый
	c.	are very interested	c.	quan tâm/say mê	c.	заинтересованный
	d.	do not like to work hard	d.	lười làm việc	d.	ленивый

Additionally, Elgort (2013) controlled for length and syntactic complexity of the four multiple choice options by translating these four options as single words or short phrases (Example 6), definitions (Example 7) or a combination of at least two single words or shorter phrases and two definitions (Example 8). In this regard, Elgort (2013, p. 259) states that a Russian translation equivalent of the tested word was provided when "at least one more out of the three remaining choices could be replaced by a Russian word. This was done to prevent guessing on the basis of length and/or syntactic complexity of the choices."

Example 6

	Eleventh 1,000 level			
	Monolingual VST item		Russian VST item	
	mussel: They bought mussels.		**mussel:** They bought mussels.	
	a.	small glass balls for playing a game	a.	бисер
√	b.	shellfish	b.	мидия **(= mussel)**
	c.	large purple fruits	c.	гранат
	d.	pieces of soft paper to keep the clothes clean when eating	d.	салфетка

Example 7

	Thirteenth 1,000 level			
	Monolingual VST item		Russian VST item	
	ubiquitous: Many weeds are ubiquitous.		**ubiquitous:** Many weeds are ubiquitous.	
	a.	are difficult of get rid of	a.	трудно выводимый
	b.	have long, strong roots	b.	с длинными крепкими корнями
√	c.	are found in most countries	c.	встречающийся повсюду **(= ubiquitous)**
	d.	die away in winter	d.	умирающий зимой

Example 8

	Tenth 1,000 level			
	Monolingual VST item		Russian VST item	
	lectern: He stood at the lectern.		**lectern:** He stood at the lectern.	
√	a.	desk to hold a book at a height for reading	a.	кафедра **(= lectern)**
	b.	table or block used for church sacrifices	b.	место жертвоприношения в храме
	c.	place where you buy drinks	c.	место, где продаются алкогольные напитки
	d.	very edge	d.	край

Nevertheless, the translation principle 2 was not followed by the Russian translation when: (1) *the tested word was an English/Russian cognate or loan word*. In this particular respect, Elgort (2013, p. 259) explains that instead of translating the cognate in the tested word, "the original choice from the monolingual VST was translated into Russian." As Example 9 illustrates, instead of using the Russian cognate (*динозавр*) of the English tested word

'*dinosaur*', a definition of what a dinosaur is was translated into Russian; and (2) *length or syntactic complexity of L1 translation of the four multiple choice options could be an indicator of the correct answer*. Note in Example 10, four well-balanced multiword noun phrases. (For further details on the translation principles used for the Russian and Vietnamese translations of the VST see Elgort, 2013; and Nguyen & Nation, 2011, respectively).

Example 9

	Third 1,000 level			
	Monolingual VST item		Russian VST item	
	dinosaur: The children were pretending to be dinosaurs.		**dinosaur:** The children were pretending to be dinosaurs.	
	a.	robbers who work at sea	a.	пират
	b.	very small creatures with human form but with wings	b.	эльф
	c.	large creatures with wings that breathe fire	c.	огнедышащий дракон
√	d.	animals that lived a long time ago	d.	ящер, живший много лет назад

Example 10

	Ninth 1,000 level			
	Monolingual VST item		Russian VST item	
	weir: We looked at the weir.		**weir**: We looked at the weir.	
	a.	person who behaves strangely	a.	тот, кто себя странно ведёт
	b.	wet, muddy place with water plants	b.	илистая местность с водяными растениями
	c.	old metal music instrument played by blowing	c.	старинный духовой инструмент
√	d.	thing built across a river to control the water	d.	сооружение на реке для контроля уровня воды

In relation to the Spanish version of the VST, it translated the same part of the multiple choice items as previous bilingual versions of the VST (Elgort, 2013; Karami, 2012; Nguyen & Nation, 2011). That is, in the bilingual Spanish VST the four multiple choice options (i.e., one correct answer and three distractors) were translated into Spanish while the tested word (i.e., target word or prompt) of each item and its non-defining context remained in English (See Appendix 1 with the Spanish bilingual version of the VST). Table 3.1 provides an example of items in the monolingual 14,000 VST and the Spanish bilingual version of this test.

Table 3.1 Example of a monolingual and bilingual Spanish VST item

Monolingual VST item (fifth 1,000 level)	Bilingual Spanish VST item (fifth 1,000 level)
HAUNT: The house is **haunted.** a. full of ornaments b. rented c. empty d. full of ghosts	HAUNT: The house is **haunted.** a. llena de adornos b. alquilada c. vacía d. llena de fantasmas

3.2.1.1 Translation principles adopted for the translation of the Spanish bilingual version of the VST

The Spanish bilingual version of the VST tries to resemble as much as possible Nation and Beglar's (2007) monolingual 14,000 version of the test. This was done mainly to preserve the design of the monolingual VST and allow comparison between the bilingual and the monolingual versions of the test. Inevitably, however, the tests are essentially different tests.

The main translation principles used for the Spanish bilingual version are as follows:

Translation principle 1: The possible answers (options) that are single words should remain as single words or simple phrases in the bilingual version. When the translated options are definitions or longer phrases, they should also be kept as definitions or longer phrases. Both cases are illustrated in Table 3.2.

Table 3.2 Example of translation principle 1

Monolingual VST (second 1,000 level)	Spanish translation (second 1,000 level)
DRIVE: He **drives** fast. a. swims b. learns c. throws balls d. uses a car	DRIVE: He **drives** fast. a. nada b. aprende c. lanza la pelota d. conduce el auto

Translation principle 2: Where possible, the two sets of options should have similar lengths to each other, because length should not be an indicator of the correct answer (See Table 3.3). That also meant that both the definitions and single words used as distractors were translated instead of being replaced by their Spanish equivalents.

Table 3.3 Example of translation principle 2

Monolingual VST (fourth 1,000 level)	Spanish translation (fourth 1,000 level)
TUMMY: Look at my **tummy**. a. cloth to cover the head b. stomach c. small furry animal d. thumb	TUMMY: Look at my **tummy**. a. trozo de tela para cubrir la cabeza b. estómago c. animal pequeño y peludo d. pulgar

Translation principle 3: The correct choice in Spanish should not contain the cognate of the tested word. For example, as shown in Table 3.4 the translation for the tested word '*period'* could have been the cognate '*periodo*', but translation principle 3 did not allow this. Another word was used so that test takers could not solely look for a similar tested form between the tested word and the correct choice.

Table 3.4 Example of translation principle 3

Monolingual VST (first 1,000 level)	Spanish translation (first 1,000 level)
PERIOD: It is a difficult **period.** a. question b. time c. thing to do d. book	PERIOD: It is a difficult **period.** a. pregunta b. tiempo c. hecho d. libro

Translation principle 4: The distractors and the correct answer should belong, as much as possible, to the same part of speech.

Table 3.5 Example of translation principle 4

Monolingual VST (second 1,000 level)	Spanish translation (second 1,000 level)
PUB: They went to the **pub**. a. place where people drink and talk b. place that looks after money c. large building with many shops d. building for swimming	PUB: They went to the **pub**. a. lugar para ir a beber y conversar b. institución en que se guarda el dinero c. edificio grande que tiene muchas tiendas d. estanque destinado a la natación

As the example in Table 3.5 shows, all the translated options are noun phrases both in the monolingual and bilingual versions of the test.

Translation principle 5: When a distractor or a correct answer has a different possible meaning, the Spanish equivalent chosen should be the one with: (1) the most frequent meaning in Spanish and, (2) the most suitable meaning for the context provided.

Table 3.6 Example of translation principle 5

Monolingual VST (fourth 1,000 level)	Spanish translation (fourth 1,000 level)
CRAB: Do you like **crabs**? a. sea creatures that walk sideways b. very thin small cakes c. tight, hard collars d. large black insects that sing at night	CRAB: Do you like **crabs**? a. criaturas marinas que caminan de lado b. galletas pequeñas y finas c. cuellos duros y ajustados d. insectos grandes de color negro que cantan en la noche

As Table 3.6 shows, the word '*collar*' in option *c* has two possible meanings in Spanish: meaning (1) a piece of clothing around the neck (i.e., '*cuello'* in Spanish), (2) a strap made of leather or other strong material which is put around the neck of an animal, especially a dog or cat (i.e., '*collar'* in Spanish). In addition, the word '*collar*' in Spanish can mean '*necklace*' in English. In the case of the English word '*collar*', the most common and adequate translation (*cuellos*) for the context provided by option *c* was selected.

Translation principle 6: In general, the translated options should make sense in the syntactic context provided by the target word. Also the single words, phrases and definitions translated should sound natural and idiomatic for Spanish speakers.

Table 3.7 Example of translation principle 6

Monolingual VST (fourth 1,000 level)	Spanish translation (fourth 1,000 level)
REMEDY: We found a good **remedy**. a. way to fix a problem b. place to eat in public c. way to prepare food d. rule about numbers	REMEDY: We found a good **remedy**. a. forma de arreglar un problema b. lugar público donde se sirven comidas c. forma de preparar comida d. regla matemática

As shown in Table 3.7, in the case of option *d* (i.e., rule about numbers) a word-by-word translation creates a Spanish phrase (i.e., regla sobre números) that is syntactically possible, but does not make much sense in Spanish. For this reason, a more idiomatic translation (i.e., regla matemática) was chosen for option *d*.

The particular Spanish translation principles explained above were developed for the purpose of this study in an effort to maintain the nature of the monolingual VST, and to produce a translation that was as true to the original test as possible and yet was still in idiomatic Spanish.

After having explained the translation principles for the Spanish bilingual version of the VST, we can notice that they differ from the principles used for the Vietnamese and Russian translations of the VST in relation to the approach followed to deal mainly with the translation of: (1) cognates, and (2) the original definitions in the four multiple choice options. That is, the translation principles used for the Spanish translation of the VST were adopted based on the impossibility to avoid the translation of cognates both in the tested words and multiple choice options, and the inability to find, in most cases, L1 single word equivalents for translating the original definitions in the four multiple choice options of the monolingual VST. Overall, the nature of the Spanish language demanded the adoption of translation principles for the bilingual English/Spanish VST in accordance with the morphological, syntactic, semantic, and pragmatic features of Spanish. This decision was also made with the intention of allowing comparisons between the monolingual and Spanish bilingual VST.

3.2.1.2 English/Spanish cognates in the VST

In relation to the vocabulary load of academic texts for speakers of Romance languages, several authors (Cobb, 2000; Nation, 1990; Read, 1988; Schmitt et al., 2001) have reported the general trend among native speakers of Romance languages (French, Italian, Portuguese, Romanian, and Spanish) to perform well in academic vocabulary tests in English. This has commonly been attributed to the fact that a high proportion of the general academic words in English – i.e., 90.70% in Coxhead's (2000) AWL – are of Greek, Latin or French origin (e.g., *period* from Greek *periodos*, *significant* from Latin *significare*, *research* from French *recherche*, *legal* from Latin *legalis* or Middle French *légal*). This high percentage of cognates – words that come from the same origin (Graeco-Latin vocabulary) and have closely related forms and meanings – may make native speakers of Romance languages have different academic vocabulary requirements from other ESP learners.

Given the fact that English and Spanish use the same letters, the Roman alphabet, the learning of the written form will be easier than for L2 learners whose mother tongues use a different script. However, English and Spanish spelling patterns are different, and Spanish speakers need to become familiar with them, even in regard to cognates.

Although English/Spanish cognates are similar in their form and meaning, their written forms have different degrees of relatedness. They range from written forms that in both languages look the same, to written cognates still related formally but with more distant morphology. To illustrate this point, Ard and Homburg's (1992) formal scale of orthographical and morphological similarities between English and Spanish has been chosen and adapted for this study. The adaptation of Ard and Homburg's scale for the present study involves the following five levels of formal relatedness.

Cognates rating scale:
0 = Identical form
1 = Similar stems with one orthographical change (one letter change)
2 = Similar stems with minor orthographical changes (two letters change)
3 = Similar stem with three orthographical changes (three letters change)
4= More distant form (four or more letters change)

Level 0: Identical form

At this level, the English and Spanish cognates look exactly the same. Examples of cognates with identical form in the tested words of the VST are *azalea, eclipse, puma, jovial, marsupial, cordillera*.

Level 1: Similar stems with a minor orthographical change

At this level, we have words whose written form is closely related to the Spanish equivalent translation, except for one letter change. Here only one orthographical change to the stem or ending of the word is required to obtain a Spanish equivalent. That is, there is only one letter being deleted or being replaced. Examples from the tested words of the VST are:

English word	Spanish equivalent
thesis	*tésis*
reptile	*reptil*

Level 2: Similar stem with two orthographical changes

At this level two orthographical changes (by adding or deleting letters) to the stem or ending of a word are required to obtain a Spanish equivalent. Examples from the tested words of the VST are:

English word	Spanish equivalent
dinosaurs	*dinosaurios*
authentic	*auténtico*

Level 3: Similar stem with three orthographical changes

Here at least three orthographical changes (by adding or deleting letters) to the stem or ending of a word are required to obtain a Spanish equivalent. Typically these changes occur in two or more different places in a word. Examples from the tested words of the VST are:

English word	Spanish equivalent
standards	*estándares*
patience	*paciencia*

Level 4: More distant form

At this level, we have cognates with four or more orthographical changes and sometimes even more than half of the letters are different and exhibit more distant orthography and morphology than at previous levels. Examples from the tested words of the VST are:

English word	Spanish equivalent
maintain	*mantener*
soldier	*soldado*

This formal scale is based on different degrees of formal similarity between English and Spanish words (see Table 3.8). In general, for Spanish speakers taking the VST it would be easier to identify the cognates at levels 1, 2 than at levels 3 or 4. That is, the closer the written form of the cognates is, the more likely they are to be identified by second language learners.

Table 3.8 Cognates rating scale

Cognates rating scale	Form	English word	Spanish equivalent
0	identical form	yoga	yoga
1	one letter change	period	periodo
2	two letters change	yoghurt	yogur
3	three letters change	peel	pelar
5	four or more letters change	compound	compuesto

Let us now consider the number of cognates, false cognates and non-cognates in the 140-item VST by using the cognates scale mentioned above to calculate the number of English/Spanish cognates among the tested words.

Table 3.9 Number of cognates, false cognates and non-cognates in the VST

140 tested words	Cognates	False cognates	Non cognates
Number of words	76	6	58
Percentage of words	54.29%	4.28%	41.43%
Examples	malign, didactic, vocabulary, deficit	figure, pro, candid, bloc, talon, gauche	drive, jump, upset, butler, quilt, haunted

As Table 3.9 shows, applying the rating scale to the 140-item VST yielded 76 cognates (54.29%), 6 false cognates (4.28%), and 58 non-cognates (41.43%) among the tested words. These figures fit nicely with the proportion of Graeco-Latin words in English found by other researchers (see Nation, 2013, pp. 390–1 for references) which is calculated as being around 60%. This suggests that the sampling of the words for the Vocabulary Size Test is not misrepresenting the language.

3.2.1.3 Checking of the translation

Once the translation criteria were set up, an experienced translator and the researcher independently translated into Spanish the four multiple-choice options of all the 140 items

of the VST. For these two parallel translations of the VST, we both followed the translation principles previously explained in section 3.2.1.1.

After the two parallel translations of all the multiple-choice options were completed, these two translations were compared and a second draft was produced and then discussed with a professional English/Spanish translator who is a proficient native Spanish speaker. Using a think aloud protocol, the professional translator went through all the draft of the Spanish translation, editing, proofreading, providing feedback on the quality of the translation, and commenting on various linguistic aspects of the translation such as word-choice, syntax, spelling, punctuation, and style.

After making the changes suggested by the professional translator, a third draft of the Spanish translation with all the correct answers highlighted (see Appendix 3.1) was sent to two different groups of four well-educated native Spanish speakers who were all proficient speakers of English.

The first group of four translation checkers were asked to complete the VST, and given the following instructions to do so: (1) put a tick next to the letter (a-d) with the closest meaning to the target word in the question, (2) pay close attention to the usage of the Spanish language while answering the test, (3) when you think there is a more appropriate Spanish equivalent, please write it down next to the Spanish word or phrase it can replace. Once again the Spanish translation was revised taking into consideration the comments of this first group of translation checkers.

Then a fourth draft of the translation was sent to the second group of four checkers who were given a version of the VST with the answer key included. This second group of translation checkers were asked to focus their attention on the appropriate usage of the Spanish language, and how adequate the translation was to their Spanish dialects. In order to achieve this, they were asked to look at word-choice, syntax, spelling, punctuation, and style. The two groups of translation checkers above mentioned came from different Spanish speaking countries, namely, Argentina, Chile, Colombia, Mexico, Peru, Spain and Venezuela. The idea of having a total of eight translation checkers from seven of the most populated Spanish speaking countries was to produce a Spanish bilingual version of the VST that could be considered as representative of standard Spanish as possible. On average

most of the changes suggested by the L1 Spanish speakers who served as translation checkers were relevant, and nine out of ten of the changes were made as a result of this checking.

3.2.2 Reorganisation of the VST items

The main reason for reorganising the multiple-choice items of the monolingual and bilingual versions of the VST was to obtain two monolingual and two bilingual VSTs with half the number of items of the 140 version of the test. In this section, the ordering scheme followed to organise the 140 items of the VST is explained. With this purpose in mind, five items from each of the fourteen 1,000 word family levels included in the test were systematically selected. This resulted in two 70-item versions of the monolingual and bilingual VST. (See Table 3.10). Likewise, this reordering scheme allowed the possibility to assign every single test taker 70 items of the monolingual version followed by 70 different items of the bilingual version, or vice versa.

Table 3.10 Reordering scheme to organise the 140 items of the VST

Group		First VST block	Items in the first VST block (from each of the 14 ten-item levels)	Second VST block	Items in the second VST block (from each of the 14 ten-item levels)
A	1	Monolingual	1,4,5,8,10	Bilingual	2,3,6,7,9
	2	Bilingual	2,3,6,7,9	Monolingual	1,4,5,8,10
B	1	Monolingual	2,3,6,7,9	Bilingual	1,4,5,8,10
	2	Bilingual	1,4,5,8,10	Monolingual	2,3,6,7,9

As summarised in Table 3.10, the 140 items of the VST were reordered in four groups (i.e., Group A.1, Group A.2, Group B.1, and Grout B.2), each group sitting the whole test but with one monolingual and one bilingual block:

- Group A.1: Monolingual VST block + Bilingual VST block: 70 monolingual multiple-choice items (items number one, four, five, eight, and ten of each 1,000 word family level) followed by 70 different bilingual items (items number two, three, six, seven, and nine).

- Group A.2: Bilingual VST block + Monolingual VST block: 70 bilingual multiple-choice items (items number two, three, six, seven, and nine) followed by 70 different monolingual items (items number one, four, five, eight, and ten of each 1,000 word family level).
- Group B.1: Monolingual VST block + Bilingual VST block: 70 monolingual multiple-choice items (items number two, three, six, seven, and nine) followed by 70 different bilingual items (items number one, four, five, eight, and ten of each 1,000 word family level).
- Group B.2: Bilingual VST block + Monolingual VST block: 70 bilingual multiple-choice items (items number one, four, five, eight, and ten of each 1,000 word family level) followed by 70 different monolingual items (items number two, three, six, seven, and nine).

Table 3.10 shows that the difference between the A groups and the B groups lies in which items were monolingual and which items were bilingual. The difference between 1 and 2 was simply an order difference. This selection of items meant that the same item in different languages was not sat by the same person.

All test takers were randomly assigned by computer to one of these four groups (i.e., Group A.1, Group A.2, Group B.1, and Group B.2), allowing comparison of the two versions and any language presentation order effects. This meant that any test taker could be assigned to any one of the four groups.

3.2.3 Participants

The tests were used with a total of 408 undergraduate L1 Spanish speakers (211 males and 197 females) who were about to start their first ESP course as part of their university degree. The average age of the participants was 21 years (Min=17, Max=25, SD=2.5). In relation to the groups to which they were randomly assigned, 209 test takers were in Groups A.1 and A.2, and 199 were in Groups B.1 and B.2. According to the language presentation order results, 207 participants sat the 70-item monolingual VST block first plus a different set of 70 items from the bilingual block, while 201 sat the bilingual Spanish 70-item VST block first immediately followed by a different set of 70 items from the monolingual block.

The students taking the VST were enrolled in the Schools of Medicine, Engineering, Economics, Humanities, Natural Sciences, and Farming Sciences of two Spanish medium higher education institutions in Venezuela. Most of the students were taking their ESP course along with other foundation courses relevant to their degrees.

Ethics clearance for the study was granted by Victoria University of Wellington Human Ethics Committee (see Appendix 5.1), and all participants gave informed consent to sit an online version of the VST.

3.2.4 Setting and administration

The specific context for this study was two Spanish medium universities in Venezuela, namely, University of Los Andes, and Polytechnic Territorial University of Mérida. In both institutions most undergraduate students take compulsory ESP reading classes during the first year of their studies.

Although the most common form of administration of this type of test for several years has been pen and paper, it was decided to use an online administration for the present study taking into consideration some of the benefits of online testing for language researchers, such as: easier randomization of items, automatic scoring of items, reaching a large number of test takers in different locations simultaneously, and the possibility of including other testing variables such as response time and immediate feedback on the final result. All the 408 test takers in this study sat an online version of the VST using a website called VocabularySize.com (http://my.vocabularysize.com/) created and maintained by Myq Larson.

The online version of the VST was administered in computer laboratories with internet access. Each computer laboratory had between 15 and 20 student computers. It took around 45 minutes for most to answer the 140-item version of the online VST. After completing the VST, the test takers received their scores (i.e., how many word families they knew in English) and a brief explanation of what that meant. The test results were saved onto an Excel spreadsheet that could be easily accessed by the language researchers or teachers administering the test at http://my.vocabularysize.com/. Since 2010 VocabularySize.com has provided language teachers and vocabulary researchers with a freely available web tool

to conduct online vocabulary testing by creating sessions, uploading tests, using automated scoring, and saving the results hosted on this website. The use of VocabularySize.com greatly facilitated the administration and analysis of the results presented in this chapter.

3.3 Results and discussion

This section answers the three research questions presented in the introduction to this chapter looking at vocabulary size, and makes a comparison between the monolingual and Spanish bilingual versions of the test.

3.3.1 What is the vocabulary size of the participants?

A total of 408 ESP students took the 140-item VST (i.e., comprising 70 monolingual and 70 bilingual items). Their mean score was 63 (*SD*=15.24) out of 140. The standard deviation and the range (maximum-minimum) are reasonably large, indicating a wide range of proficiency among the participants. (See Table 3.11).

Table 3.11 Results of the 140-item VST (comprising the monolingual and bilingual test results together)

14,000 VST	Number of test takers	Minimum score	First quartile	Mean score	Third quartile	Maximum score	SD
140 items	408	29	52	63	73	109	15.24

The vocabulary size of the students taking the 140-item VST was calculated by multiplying the total number of correct answers by 100, because each word in the test represents 100 items in the BNC word family lists. As summarised in Table 3.11, the scores of the test takers ranged between a minimum score of 29 (i.e., 2,900 word families), and a maximum score of 109 (i.e., 10,900 word families). The range of scores means that the vocabulary size of the test takers in this study ranged from low to intermediate receptive lexical proficiency levels.

Based on the results of the VST in Table 3.11, L1 Spanish speakers know an average of 6,300 word families at the start of their first ESP course, and have an intermediate receptive vocabulary proficiency level in English.

3.3.2 Is the bilingual version of the VST a better indicator of vocabulary size for L1 Spanish speakers?

To examine the differences between taking the monolingual or the bilingual VST, the total scores of each 70-item version of the test were compared separately.

Table 3.12 Monolingual versus bilingual VST results

VST version	Minimum score	First quartile	Mean score	Third quartile	Maximum score
Monolingual (70 items)	11	25	31.40	37	55
Bilingual (70 items)	12	25	31.60	38	54

The results in Table 3.12 demonstrate the minimum scores (11 correct answers vs. 12 correct answers) and the maximum scores (55 correct answers vs. 54 correct answers) of the monolingual and bilingual VST differed only by one correct answer. These results also revealed that the participants obtained only a slightly higher mean score (31.60) in the bilingual VST.

To obtain the average number of word families known by the participants, mean scores were multiplied by 200 (Elgort, 2013). When we looked at the mean scores (i.e., 31.40 = 6,280 word families vs. 31.60 = 6,320 word families) of the monolingual and bilingual VST respectively, the test takers scored an average of 40 more word families in the bilingual VST.

In order to determine whether the differences in mean scores presented in Table 3.12 were significant, a series of pair-wise comparisons were conducted. First, a Welch two sample t-test was carried out to compare the mean scores between the monolingual and the bilingual VST. The results indicated that there was no significant difference in using the two versions, $t(814)=0.29$, $p=0.76$.

Despite the fact there was no significant difference between taking the bilingual or monolingual VST, we observed in the testing room that test takers seemed more willing to take the bilingual Spanish VST and looked more relaxed. In general, the test takers' anxiety seemed to be lower when taking the bilingual VST.

Some of the comments the test administrators could hear the test takers constantly make while talking among themselves after taking the VST were as follows:

- "It was easier for you to get your head around what you are doing when you take the bilingual VST."
- "It takes more time to find the correct answer when you take the monolingual VST."
- "You feel more tired when you take the monolingual VST."

Based on some of the comments made and the attitudes observed in the testing room, we decided to further explore whether there was a fatigue effect involved while taking the VST (See Table 3.13).

Table 3.13 Fatigue effect in the VST

VST version (presentation order)	Minimum score	Maximum score	Mean score
Monolingual VST (first)	14	55	35.00
Bilingual VST (first)	14	54	34.90
Monolingual VST (second)	11	50	27.70
Bilingual VST (second)	12	54	28.40

In order to determine if the differences observed in the scores for the items presented in the first and second blocks of the VST, the scores in each language for the first block and the second block were compared. The results in Table 3.13 indicate that the scores for the first block (35.00 and 34.90) were always higher than the scores for the second block (27.70 and 28.40), which were always lower both for the English version, $t(406)=9.12$, $p<001$ and Spanish version, $t(400)=7.89$, $p<001$.

When we look at the order in which both VST versions were presented, we observe that the differences summarised in Table 3.13 between taken the first block and the second block of the test were not significant between languages (see Table 3.14), which suggests that the test takers were tired after taking the first 70 items of the VST no matter which version of the test they sat first.

Table 3.14 Results by block including both VST versions

VST version (presentation block)	Minimum score	Maximum score	Mean score	SD
Monolingual and bilingual (first)	14	55	34.95	7.97
Monolingual and bilingual (second)	11	54	28.03	8.46

As shown in Table 3.14, the differences between the scores for the first block were not significant, $t(405)=0.17$, $p=0.8$. Likewise, no significant differences were found between the scores for the second block either, $t(403)=0.85$, $p=0.39$. The lower scores in the second block were probably caused by a fatigue effect.

3.3.3 Is there a difference in scores as a result of the language presentation order?

To investigate whether there was a significant difference between taking the monolingual or the bilingual block of the VST first, the results of each block of the test were compared separately. The results of taking the 140-item VST showed that there was no order effect caused by starting with the bilingual or the monolingual VST (See Table 3.15).

Table 3.15 Presentation order of the monolingual and bilingual VST

VST blocks	Minimum score	Maximum score	Mean score
Monolingual block first	14	55	35.00
Bilingual block first	14	54	34.90

As shown in Table 3.15, the group of test takers sitting the monolingual block first only gained a slightly higher maximum score of 55 correct answers (i.e., 11,000 word families) than the group taking the bilingual block first (54 correct answers, i.e., 10,800 word families). The minimum scores were the same (Min=14, i.e., 2,800 word families) for both VST versions.

Since the mean scores (35.0 vs. 34.9) of both language blocks were very similar, a t-test was used to determine whether there was a significant difference depending on which block was sat first (i.e., the monolingual or the bilingual block). The results of the t-test comparing the scores of two versions of the VST revealed that the differences were not significant, $t(814)=0.56$, $p= 0.57$.

Overall, the results previously summarised on the administration of the VST showed the following: (1) the average vocabulary size of L1 Spanish university students about to start their first ESP course is 6,300 word families; (2) there was no significant difference between taking the bilingual Spanish or monolingual VST first or second; (3) both versions of the VST were equal measures of the vocabulary size of the participants; (4) there was a fatigue effect (regardless of language) observed after taking the first 70-item version of the VST.

3.4 Limitations

The limitations of the VST include the following factors:

1. *The multiple choice format* which allows guessing either through a process of elimination among the choices (Gyllstad, Vilkaite, & Schmitt, in press), or through random choice. This may result in an overestimation (Gyllstad et al., in press; Stewart, 2014) of the actual vocabulary size of the test takers. Zhang (2013) shows that the inclusion of a fifth option (i.e., the I don't know option) discourages blind guessing and helps provide a slightly more reliable estimate of the learner's vocabulary size. In this regard, however, Nation (2012) notes that the 'I don't know option' prevents test takers from drawing on subconscious/partial knowledge of the tested words when making informed guesses of the correct answer.
2. *Criterion for the selection of the tested words.* Instead of adopting mainly a frequency-based criterion for the selection of the tested words, other word selection criteria such as word difficulty (Beglar, 2010) and proportion of cognates (Elgort, 2013) between the L2 (i.e. English) and the L1 of the test takers could also be taken into consideration as additional criteria. In relation to the cognate criterion, Elgort and Coxhead (in press) draw attention to the importance of controlling for the number of cognates in bilingual versions of the VST created for particular groups of test takers who are non-native speakers of English. This could lead, according to Elgort and Coxhead, to more accurate estimates of the vocabulary size of a given group of test takers who share the same L1.
3. *Ranking of the items in the test* which instead of presenting the tested words in progressive descending frequency order (from higher to lower frequency levels) which was the case for the administration of the 70-item versions of the original

14,000 VST already described in this chapter, could spread the tested words with high, mid, and lower frequencies through the whole test. By mixing the frequency order of the tested words appearing in the test, the test takers are more prone to remain interested and involved in the test (Nguyen & Nation, 2011), and the time they spend answering the test remains, on average, more stable for all the tested items (Nation & Coxhead, 2014).

Some of the limitations of the results of the VST reported in this chapter are related to the administration format used. Despite the undeniable advantage of administering the test online, it also meant in some cases: (1) limiting the administration to classes with internet access and computer facilities, and (2) only being able to administer the test simultaneously to a smaller number of participants than the actual number of students willing to take the test (capacity in computer labs ranged from 15 to 20 computers with internet connection). Nevertheless, an online administration format was chosen to facilitate gathering, processing, and analysis of the results. The online test format also gave test takers the opportunity to know their test scores as soon as they finished answering the test.

3.5 Conclusions and implications for the study

In this chapter we looked at the value of measuring the vocabulary size of potential students of introductory ESP courses; then we explained the translation principles applied to the bilingual Spanish VST and the role played by cognates in the translation process; finally we discussed the differences between administering a monolingual versus a bilingual VST to L1 Spanish learners of English as a foreign language.

The main reason for identifying the vocabulary size of the group of prospective ESP students was so that we could investigate the gap between the number of words they knew and the number of words they needed to learn in order to be able to understand academic texts written in English. However, the identification of the learners' vocabulary size may also serve a variety of other important purposes in the ESP classroom. For instance, an identification of the vocabulary size of these ESP learners could be used to make vocabulary lists that meet the vocabulary needs of this group of ESP learners, develop vocabulary tests that determine the degree to which an ESP course or programme is really meeting its lexical objectives, and allow ESP educators to plan the vocabulary component

of these ESP courses in a more informed manner. Knowing learners' vocabulary sizes also allows individual remedial work to target those in need.

No identification of the English vocabulary size of native Spanish speakers who are prospective learners of ESP has yet been published, despite the fact that such identification could serve a number of important roles (i.e., for course placement, for research purposes, for measuring vocabulary growth, and for guiding the design of suitable course materials, etc.) in the teaching and learning of vocabulary. The results of this chapter provide a useful measure of the vocabulary size of this particular group of ESP learners. These results will be used later on in Chapters 4, 5, and 6 as a useful point of reference for investigating the relationship between vocabulary size and vocabulary load of medical texts. The results of the VST here presented could well be applied to the investigation of the relationship of vocabulary size and vocabulary load of other subject-specific areas, and academic texts that will not be investigated in this study.

After presenting the results of the administration of the VST to a group of native Spanish speakers in Venezuela, we hope the vocabulary size scores previously discussed in this chapter may have shed some light on the vocabulary load of academic texts for ESP learners whose first language (Spanish) has many words of Latin and Greek origin. The results here presented will also help us get a more realistic picture of the relationship between vocabulary size and vocabulary load of academic texts in English for native speakers of Romance languages (native Spanish speakers, in particular). Moreover, these results will be useful to establish in subsequent chapters of this thesis the relationship between the average vocabulary size and the number of words ESP students need in order to be able to understand medical texts written in English.

Chapter 4: Medical vocabulary and the GSL, the AWL, and the Pilot Science List

4.1 Introduction and justification

In this chapter, a traditional approach for estimating the number of words needed to achieve an appropriate lexical coverage (i.e., 98%) of medical textbooks has been adopted. This traditional approach has specifically led to the use of popular general (e.g., the GSL by West, 1953) and academic (e.g., the AWL by Coxhead, 2000) word lists plus the creation of new subject-specific (medical) word lists beyond these well-known existing word lists (i.e., GSL and AWL) to reach a good lexical threshold of specialised texts, i.e., medical textbooks in the particular case of this study. As we have mentioned above, the traditional approach to the selection of words that should be prioritised in ESP and EAP courses has involved the use of existing word lists - in particular, West's (1953) GSL, Coxhead's (2000) AWL, and more recently Coxhead and Hirsh's (2007) Pilot Science List. In addition, this approach has required the development and lexical analysis of specialised corpora in a wide range of studies (Chen & Ge, 2007; Chung & Nation, 2003, 2004; Coxhead et al., 2010; Fraser, 2007; Hsu, 2013, 2014; Konstantakis, 2007, 2010; Sutarsyah et al., 1994; Wang et al., 2008; Ward, 1999, 2009, among others) already discussed in Chapter 2. Operating within this framework, the present investigation developed medical (technical) word lists which incorporated these widely used existing word lists (i.e., GSL, AWL and Pilot Science List).

A similar corpus-based approach to the one followed by the aforementioned researchers was adopted in this chapter to investigate the number of English words (the vocabulary load) that ESP medical students need to know in order to reach an optimal lexical threshold (i.e., 98% of the tokens in a text) and thus be able to access a wide variety of medical texts written in English.

The main goal of this chapter is to test this traditional approach to examining the vocabulary load of medical textbooks. The first big step is to decide on a medical corpus for the study and then prepare it for analysis. The second step was to prepare a general corpus of the same size so that the corpus comparison method could be used. The third step involved the

regularisation of the existing word family lists (i.e., in the GSL, AWL, and Pilot Science List) to make sure that the word families of the same words used in the various lists were exactly the same. The fourth step of the corpora and word list preparation process for this study consisted of devising a system for identifying medical words which was based on Chung and Nation's (2003) rating scale for classifying specialised words. The fifth step was to develop a series of medical technical word lists through corpus comparison.

Once these steps are completed, the results of running the GSL, AWL, Pilot Science List, and the new medical lists through the medical corpus are presented. This is followed by a discussion of the behaviour of the various lists and the words not found in any of the word lists used. After that, the lexical profiles of the medical and general corpora are compared. This chapter concludes by reflecting on some of the instructional implications of aiming for 98% lexical coverage of medical texts.

4.2 Compilation of the medical corpus

In order to examine the vocabulary load of medical textbooks written in English, a medical corpus was compiled for this study. Before compiling the medical corpus, an ESP university teacher with over 25 years of experience in teaching and researching medical English at undergraduate and graduate level was consulted on the most suitable texts and topics to be included in the medical corpus.

With first and second year medical students in mind, she suggested textbooks like *Harrison's Principles of Internal Medicine,* 17th edition (Fauci et al., 2008), and *CECIL Medicine*, 23rd edition (Goldman & Ausiello, 2008), which include a comprehensive range of medical topics. These textbooks are used from the first year of medical studies. She also noted that these textbooks are widely consulted by undergraduate and graduate medical students, and health professionals. Likewise, the group of medical students who took the VST (see Chapter 3) use these two medical textbooks from the start of their degrees.

Based on this ESP teacher's recommendations, these two medical textbooks mentioned above were chosen for use in compiling the medical corpus (i.e., the med1 corpus) for this study. The specialised texts included in the two medical textbooks used to create the med1 corpus for the present study were downloaded from various publishers and databases such

as McGraw-Hill, Elsevier, and ScienceDirect, made available by Victoria University of Wellington for research purposes. Table 4.1 contains a description of the main features of these two textbooks.

Table 4.1 Features of the two medical textbooks

Features	Harrison	CECIL
Length	2,866,130 tokens	2,565,610 tokens
Number of chapters	431 written by different authors	467 written by different authors
English variety	American English	American English
Genre	Written academic English	Written academic English

Table 4.1 shows that at over 2½ million words long, these books are very substantial texts. They are much longer than hard-copy texts in most other disciplines which are around 300,000 words long (Sutarsyah et al., 1994). They provide a very wide coverage of medical topics (e.g., abnormalities, addictions, body organs, care setting, chemicals, complications, conditions, consultation, critical care, curative substances, cures, diagnosis, disabilities, diseases, disorders, dysfunctions, healing process, health care, health management, health practices, health professionals, human body, injuries, laboratory values, living organisms, measurements, medications, minerals, nutrients, nutrition, organ systems, pain management, patients, recovery, remedies, surgeries, symptoms, syndromes, tests and procedures, therapies, transplantation strategies, treatments, and vitamins, among others).

The compilation procedure, and the inclusion and exclusion criteria followed to create a medical corpus of 5,431,740 tokens are summarised in Table 4.2. The decision to delete the items mentioned in the exclusion criteria in Table 4.2 was made based on results from piloting using a 500,000 token medical corpus compiled from a balanced sample of text extracts from *Harrison's Principles of Internal Medicine* (Fauci et al., 2008). The pilot analysis showed a great number of proper names, abbreviations from bibliographies and URL addresses with proper names, and abbreviations that in the medical corpus have specialised meanings which differ from other uses of the same abbreviations. The proper names found in the medical texts in this study normally referred to diseases, medical conditions, abbreviations and acronyms very common in medical texts but not so common in other kinds of texts (e.g., Crohn in Crohn's disease, HUS in Hemolytic Uremic

Syndrome, IL in interleukin, NASH in nonalcoholic steatohepatitis, and Wegener in Wegener's granulomatosis, among others).

Table 4.2 Compilation of the two medical textbooks

Compilation	Harrison	CECIL
Compilation procedure	Convert the digital version (chm files) to plain text, and delete the information listed in the exclusion criteria.	Convert the digital version (chm files) to plain text, and delete the information listed in the exclusion criteria.
Inclusion criteria	392 chapters 39 electronic chapters 1 Appendix	467 chapters 1 appendix 1 guide to the approach of common symptoms, signs, and laboratory abnormalities.
Exclusion criteria	Cover, preface, copyright information, list of contributors, acknowledgements, bibliography, further readings, all bibliographical references and acknowledgements below tables and figures, names of authors, URL addresses.	Cover, front matter, associate editors, preface, contributors, global advisory board, features of the 23rd edition, references, suggested readings, all bibliographical references and acknowledgements below tables and figures, names of authors, URL addresses.

For this main study, a med1 corpus of 5,431,740 tokens consisting of the two medical textbooks mentioned in Table 4.1 and 4.2 was used. This allowed the inclusion of a wide variety of medical topics, the use of complete texts, and a large enough coverage of topics to provide a representative sample of medical lexis.

4.3 Compilation of the general corpus

A corpus of general English was compiled to serve as a comparison corpus for this study. The general comparison corpus was built using substantial sections of several well-known English corpora, namely, the Freiburg-LOB corpus (FLOB corpus; Mair & Ludwigs, 1999b), the Freiburg-Brown corpus (FROWN corpus; Mair & Ludwigs, 1999a), the KOLHAPUR corpus (Shastri, 1986), the Lancaster-Oslo/Bergen corpus (LOB corpus; Leech, Johansson, & Hofland, 1978), the Wellington corpus of written New Zealand English (WWC; Bauer, 1993), a standard corpus of present-day edited American English (BROWN corpus; Kucera & Francis, 1964), and the Australian corpus of English (ACE; Peters, Collins, & Blair, 1986). Table 4.3 contains a description of the main features and

compilation criteria of these seven general corpora. See Table 4.3 below with more details of the general corpora.

Table 4.3 Features and compilation criteria of the general corpus

Features	General comparison corpus
Total length	5,431,740 tokens
Number of corpora	7 general corpora
Number of running words (tokens)	FLOB corpus (859,823 tokens) FROWN corpus (862,516 tokens) KOLHAPUR corpus (851,678 tokens) LOB corpus (857,255 tokens) WWC corpus (856,274 tokens) BROWN corpus (857,134 tokens) ACE corpus (287,060 tokens)
English variety	FLOB corpus (British English) FROWN corpus (American English) KOLHAPUR corpus (Indian English) LOB corpus (British English) WWC corpus (New Zealand English) BROWN corpus (American English) ACE corpus (Australian English)
Genre	Written general English
Inclusion criteria	All general English sections
Exclusion criteria	Learned section (section J)
Compilation procedure	Delete the learned and scientific section from the 7 general corpora used, join the remainder of the 7 corpora into one text file.

The general corpus was the same size as the medical corpus. This facilitated calculation of the ratio of the frequencies and avoided distortion from adjusting for different corpus sizes.

The learned section (section J) including general academic texts was removed from all the general corpora used before compiling them into a comparison corpus for this study. This was to ensure that there were no overlapping topics in the two corpora which could upset frequency comparisons. If the general corpus contained medical texts, this would confound frequency comparisons with the medical technical words.

The ACE corpus (Peters et al., 1986) used has around 500,000 fewer tokens than the other general corpora used for this study, because after removing the learned section of the ACE corpus (Peters et al., 1986) only the sections needed to reach the 287,060 tokens required to compile a 5,431,740 token general comparison corpus (the same size as the medical corpus) were included. Even though, both corpora (i.e., the general and medical corpora) had the same number of running words (i.e., 5,431,740 tokens), the texts in the medical corpus varied more in length while all the individual texts in the general corpus had an average of 2,000 words. The seven general corpora mentioned above were chosen to create the general comparison corpus for the present study, because these are well-known general corpora that have been commonly used in previous investigations (Chung & Nation, 2003, 2004; Sutarsyah et al., 1994) that have also followed a corpus comparison approach to look at the lexical profile of specialised texts of particular subject areas (i.e., anatomy, applied linguistics, and economics).

4.4 Preparing the GSL, AWL, and Pilot Science List

Taking a traditional approach to analysing the vocabulary load requires using existing word lists. Because this thesis also uses the BNC/COCA lists (see Chapter 5) in addition to the longer established GSL, AWL, and Pilot Science List (the traditional lists), it was necessary to make sure that the word families for the same words used in the various lists followed the same criteria. When comparing the BNC/COCA lists with the other lists, it was found that the division of words into families and the members of those families were not exactly the same. It was important that the GSL, AWL, and Pilot Science List should use the same word families as those used in the BNC/COCA lists, because otherwise comparisons between the traditional lists and the BNC/COCA lists would be confusing. This standardising of the traditional lists meant regularising what is included in a word family. This resulted, for example, in the GSL word family *defend* (with six word family members) becoming two different word families *defend* (with one new, *undefended*, and one excluded, *defence*, word family member) and *defence* (with twelve new word family members: *defenceless, defencelessly, defences, defense, defenses, defensible, defensibly, defensive, defensively, defensiveness, indefensible, indefensibly*). This also meant the word family *administrate* in the AWL became three different families, with the old family members *administrate* and *administrates* not appearing in the updated list, because these

words did not occur in the academic corpus and thus would not get in to the AWL. For an example of the changes to the *administrate* word family see Table 4.4.

Table 4.4 Changes to the *administrate* word family

AWL	Regularised AWL
administrate administrates administration administrations administrative administratively administrator administrators	administration administrations admin administrative administratively administrator administrators

Table 4.5 summarises the number of changes to the headwords and word family members of the GSL, AWL, and Pilot Science List.

Table 4.5 Comparison of the words in the original GSL, AWL, and Pilot Science List with the regularised versions

Lists	N° word families	N° word types
1st thousand GSL	998	4119
Regularised 1st thousand GSL	1132	7213
2nd thousand GSL	988	3708
Regularised 2nd thousand GSL	1036	6238
AWL	570	3107
Regularised AWL	600	4011
Pilot Science List	318	2015
Regularised Pilot Science List	317	2005

As Table 4.5 shows, the regularisation resulted in an increase in the number of headwords and family members in the regularised GSL and the AWL. This increase is due to the addition of family members to the already existing word families found in the GSL and the AWL, and to the creation of two or more families from a previously single family. All the new words added to the GSL and AWL meet the word family inclusion criteria up to level 6 proposed by Bauer and Nation (1993). The Pilot Science List had a slight decrease in the number of family members and headwords as a result of careful editing some minor inconsistencies.

At first, the changes to the AWL resulted in an AWL with 681 word families compared to the original 570. The 111 new AWL word families were then run through Coxhead's academic corpus (cited in Coxhead, 2000) to confirm whether these new word families met the frequency and range inclusion criteria established by Coxhead (2000, p. 221) when first making the AWL. These criteria were:

1. No families or family members were in the GSL.
2. Each word family needed to occur at least 100 times in the academic corpus.
3. Each word family needed to occur at least 10 times in each of the four academic subsections of the same academic corpus.

Thirty-eight of the 121 potential new headwords met these three criteria.

In addition to these 38 new word families, three word families that occurred at least 10 times in each of the four academic sub-sections and at least 90 times in Coxhead's academic corpus (cited in Coxhead, 2000) were also included. In spite of occurring fewer than 100 times (i.e., the word families *conversion* and *ethical* occur 90 times, and *conversely* occurs 99 times) in the academic corpus, these three additional word families were included in the updated AWL because a minimum total word family frequency of 80 occurrences was a criterion also used by Coxhead (2000) for a few word families with only one family member (*forthcoming* was the only one-member word family occurring less frequently), and also because all the word types of these three new word families occurred in Coxhead's (2000) original AWL.

In the case of the Pilot Science List the word family *defense* was deleted to avoid overlap with a similar existing word family in the second 1,000 GSL.

With all these changes, the GSL, AWL and Pilot Science List all use non-overlapping word families and families with exactly the same word family members as those used in the BNC/COCA lists. This allows sensible comparisons between the coverage of the lists.

4.5 Development of a system for identifying medical words

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. 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. These potential technical words then need to be checked systematically to decide if they are 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 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 (Step 1 of Chung and Nation's rating scale) and (2) content area (technical) vocabulary (Steps 2, 3, and 4 of the same scale).

Meaning is the main distinctive 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.

1. *General purpose vocabulary* refers to the words needed to write or talk about a wide range of topics, disciplines or content areas. Chung and Nation (2003) defined them as words not semantically related to the particular specialised field they were investigating (anatomy). General purpose vocabulary in academic texts includes

common high-frequency vocabulary found in general English word lists like the GSL and AWL, and can also include mid-frequency and low-frequency words. The meaning of these words is not closely related to the particular subject area being investigated. For example, words like *is, consequence, because, outside, ignore* can be considered as general purpose words in medical texts.

2. *Content area (technical) vocabulary* refers to words whose meaning is related to a particular topic, discipline, field or subject domain. This content area vocabulary can include high, mid and low-frequency words. We can distinguish four sub-levels of such content area words related to the medical field. According to the degrees of technicality exhibited by the medical words, Table 4.6 describes these four sub-levels of content area words in medical texts.

Table 4.6 Sub-levels of content area words in medical texts

Sub-level 2.1 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 2.2 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 2.3 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 2.4 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.

The previous criteria were used to decide whether the words from the GSL, AWL, Pilot Science List and medical word lists were general or medical words. Manual checking was used to check 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 (such as the online dictionary http://medical-dictionary.thefreedictionary.com/), and (2) confirming the medical senses of these in their actual context of occurrence in the med1 corpus (i.e., the medical corpus originally compiled for creating the medical word lists).

During the checking, two problems were encountered.

Problem 1: Abbreviations

Some abbreviations in medical English are counted as members of general word families by Range (Heatley et al., 2002). Some of these abbreviations appear normally in uppercase (DR, DAD, PET, DRESS, FISH) or have a combination of uppercase and lowercase letters (Dr, DADs, PETs, AGEs, HbsAg, *hCG, pH, mOsm, SpAs, pANCAS, PTHrP, HbS, PaCO2, NaCl, q8h, qd, mL, E31*). Also the meaning of these abbreviations is different from the meaning in general usage. Here are some examples. The word type DADs is in the first thousand GSL, but it means '*delayed after depolarisations*' in the medical corpus. The word type PET is in the second thousand of the GSL but it means '*Positron emission tomography*' in the medical corpus.

Probably the best way to avoid the overlap of homographs with different uppercase and lowercase letter combinations would be to make the Range software (Heatley et al., 2002) case sensitive. If a case sensitive feature were added to Range the making of a list of common abbreviations and acronyms in medical texts written in English would be more precise and less time-consuming. Such a change however would create problems because the word families used would need to include capitalised forms such as *Dad, Pet, Pets*.

Problem 2: Word families and word types with two or more meanings

The initial classification of the technical words was based on word types rather than word families. This is because, as the examples below show, even though the GSL, AWL and Pilot 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 Pilot 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 illustrated in Table 4.7. The meanings of the words presented in Table 4.7 were looked up in the general (http://www.thefreedictionary.com/dictionary.htm) and medical (http://medical-dictionary.thefreedictionary.com/) sections of TheFreeDictionary's website.

Table 4.7 Examples of word types in existing lists with technical meanings

	GSL, AWL meaning	Medical corpus meaning
1st thousand GSL	FILLING (verb) = make full	*FILLING (verb, noun) = make full, material
2nd thousand GSL	CULTURE (noun) = beliefs	*CULTURE (noun) = growing microorganisms
2nt thousand GSL	PATIENT (adjective) = easy-going, tranquil	*PATIENT (noun) = health care recipient
AWL	RADICAL (adjective) = extreme	*RADICAL (noun) = an atom
AWL	MEDIA (noun) = means of mass communication	*MEDIA (noun) = substances used to cultivate living cells

The semantic ambiguity that might be caused by abbreviations and homographs illustrates two of the problems experienced while identifying medical words. Ideally, an inter-rater reliability check (as reported by Chung & Nation, 2003; Fraser, 2005) should have been carried out to confirm reliable use of the rating scale. Nevertheless, a single rater (the researcher) was involved in classifying words for the present study due to the impossibility of logistically having another rater (a medical specialist who is a fluent English speaker) involved in categorising the data. In spite of the fact that a single rater was used to classify the medical words for the study, the thorough classification process adopted for identifying medical words is a good indication of the reliability of the medical word lists used for the present investigation.

The steps we have looked at so far include compiling and preparing the corpora, preparing and regularising the existing word lists, and developing a checking system for making sure

the words are technical words. The next step was to carry out the corpus comparison procedure to find potential technical words.

4.6 Developing a medical technical word list through corpus comparison

The corpus comparison procedure involved largely following Chung's (2003) procedure to find potential technical vocabulary. Corpus comparison involves using a non-technical corpus and a technical corpus to compare word frequencies. Words occurring only in the technical corpus or with a much higher frequency in the technical corpus have a very high likelihood of being technical words.

For this first study, the Range software (Heatley et al., 2002) was used to carry out the frequency comparison of the medical and the general corpora. The first thousand (GSL1) and second thousand (GSL2) of the GSL, AWL, and Pilot Science List were used because the words in these lists are assumed to already be known. 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.

4.6.1 Comparing frequencies in the two corpora

This first step of the procedure involved creating a frequency list with the word types found in both the medical and general corpora. This frequency list file was then sorted by relative frequency in the two corpora. For example, the word *syndrome* occurred 7,978 times in the medical corpus and 20 times in the general corpus, giving a relative frequency of 398.90. The relative frequency figure was calculated by dividing the frequency figure of a word type in the medical corpus by the frequency figure of the same word type in the general corpus. Once this relative frequency figure was estimated for all the word types occurring in both the medical and general corpora, the word types with the highest relative frequency figures were checked using, as outlined in the previous section (4.5), the yes/no decision-making to see if they are medical terms using the definitions presented in Table 4.6. This resulted in a medical word list consisting of 3,000 word types.

The following is a more detailed description of the steps taken to create the medical word lists through relative frequency.

Step one: Inserting into a spreadsheet file all the words in both the medical and general corpora not found in the GSL, AWL, and Pilot Science List which occur with a frequency of 1 or more.

Step two: Sorting the medical and general word types not found in the GSL, AWL, and Pilot Science List by relative frequency, with the word types with the highest relative frequency at the top of the list. This means that words occurring most frequently in the medical corpus have a high positive relative frequency figure. Table 4.8 contains the top twelve medical words according to their relative frequency, which is estimated by dividing the frequency of a word type in the medical corpus by the frequency of the same word type in the general corpus.

Table 4.8 The top ten medical words organised by the higher relative frequency (Fx) (MedcorpusFx/GencorpusFx) in the medical corpus

Rank	Word type	MedcorpusFx	GencorpusFx	MedcorpusFx/GencorpusFx
1	vascular	2463	2	1231.50
2	viral	2240	2	1120
3	lesions	4795	5	959
4	lesion	1653	2	826.50
5	meningitis	1528	2	764
6	DNA	1952	3	650.67
7	gastrointestinal	1853	3	617.67
8	ventricular	2415	4	603.75
9	atrial	1182	2	591
10	CT	1743	3	581
11	mutations	2798	5	559.60
12	antibiotic	1114	2	557

Step three: Deciding which word types from the list have a medical meaning using the adaptation of Chung and Nation's (2003) semantic rating scale.

Step four: Selecting only the words that have been classified as medical words and have been sorted by highest relative frequency to make the medical word type list. These steps resulted in three new medical word lists with 1,000 word types each. These three medical lists were sorted with the highest relative frequency at the top of the lists and having a frequency equal or greater than thirteen in the medical corpus. Once there were no more word types sorted by relative frequency that provided good coverage over medical texts, a second medical word selection procedure was put into practice to continue looking for more medical word types needed to achieve 98% lexical text coverage over medical written texts, as described below.

4.6.2 Checking the words unique to the medical corpus

The words occurring only in the medical corpus and not at all in the general corpus are highly likely to be technical terms in medicine (Chung, 2003). These words unique to the medical corpus need to be checked however to make sure they really are technical terms and not just low-frequency words. This second word selection procedure also followed four steps.

Step one: Making a list of the word types occurring only in the medical corpus.

Step two: Sorting the list of words occurring only in the medical corpus by frequency, placing the word types with the highest frequency of occurrence at the top of the list.

Step three: Checking which word types from this list of unique medical words have a medical meaning.

Step four: Dividing up the medical word types into lists of 1,000 word types each. This resulted in 23 lists of 1,000 word types each.

The two main procedures used to develop new medical words lists by (1) comparing frequencies in the medical and general corpora (see Appendix 2.2 with the first 1,000

medical words occurring both in the medical and general corpora) and (2) identifying words unique to the medical corpus (see Appendix 2.3 with the most frequent 1,000 medical words occurring only in the medical corpus) were explained and illustrated in this section. Next, the steps followed to divide the medical words into new specialised word lists are presented.

4.7 Division of the medical words into word lists

The classification of the medical word types was done by: a) choosing the most frequent 3,000 medical word types occurring in both the medical and general corpora (word lists MGEN 1st 1,000, MGEN 2nd 1,000 and MGEN 3rd 1,000) with frequencies in the medical corpus ranging from 5001 (e.g. pulmonary) to 13 (e.g. enuresis) 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 b) selecting 23,000 unique medical word types (word lists MED 1st 1,000 to MED 23rd 1,000) with frequencies in the medical corpus ranging from 5,816 (e.g., renal) to 1 (e.g. dephosphorylates). From word list MGEN 1st 1,000 onwards only word types that have been previously classified as medical words using the yes/no decision-making procedure were included in the new 26 medical word lists. Table 4.9 shows the details of the three 1,000 MED word lists.

Table 4.9 The three lists totalling 3,000 word types occurring in both the medical and general corpora organised by relative frequency (Fx)

Word list	Number of word types	Range of frequencies	Relative frequency range (medical Fx/general Fx)	Example
MGEN 1st 1,000	1000	5001 to 22	5001 to 21.75	syndromes
MGEN 2nd 1,000	1000	9370 to 15	21.67 to 4	radiologist
MGEN 3rd 1,000	1000	4503 to 13	3.95 to 0.01	anatomical

Table 4.10 below gives the details of the twenty-three frequency-ranked 1,000 MED word lists that were unique to the medical corpus. As can be observed in Table 4.10 the cut off at 1,000 medical word types for each of the 23 MED lists involves the inclusion – in most of the 23 MED lists – of different medical word types in different MED lists with the same frequency of occurrence (e.g., *subcutaneously* in MED 1st 1,000 and *polyarteritis* MED 2nd 1,000 both with 111 occurrences, *proptosis* in MED 2nd 1,000 and *catarrhalis* in MED 3rd

1,000 both with 56 occurrences, and so forth). Note also in Table 4.10 the quick drop in range of frequencies particularly from MED 1st 1,000 to MED 2nd 1,000.

The word lists in Table 4.10 show the very large numbers of words making up a medical vocabulary. 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. A reason for this is that these two kinds of words may involve different learning procedures. Moreover, ranking the two types of lists separately provides better coverage with a smaller amount of word types than ranking them together by overall frequency and range. (See Chapter 7 for further discussion).

Table 4.10 The 23 MED lists of the words unique to the medical corpus

Word list	Number of word types	Range of frequencies	Examples
MED 1st 1,000	1000	5816 - 111	retinitis (135)
MED 2nd 1,000	1000	111 - 56	urticarial (113)
MED 3rd 1,000	1000	56 - 36	caries (38)
MED 4th 1,000	1000	36 - 25	lacrimal (27)
MED 5th 1,000	1000	25 - 18	angiopathy (24)
MED 6th 1,000	1000	18 - 14	toxemia (16)
MED 7th 1,000	1000	14 - 11	angioma (11)
MED 8th 1,000	1000	11 - 9	fibular (10)
MED 9th 1,000	1000	9 - 7	tuberosity (8)
MED 10th 1,000	1000	7 - 6	extramuscular (6)
MED 11th 1,000	1000	6 - 5	coagulative (5)
MED 12th 1,000	1000	5 - 4	angioscopy (4)
MED 13th 1,000	1000	4	neurovirulence (4)
MED 14th 1,000	1000	4 - 3	cecostomy (3)
MED 15th 1,000	1000	3	necrosum (3)
MED 16th 1,000	1000	3 - 2	abasia (2)
MED 17th 1,000	1000	2	dentine (2)
MED 18th 1,000	1000	2	infrarenal (2)
MED 19th 1,000	1000	2	oxaprozin (2)
MED 20th 1,000	1000	2	seromas (2)
MED 21st 1,000	1000	2 - 1	xylocaine (2)
MED 22nd1,000	1000	1	callosus (1)
MED 23rd 1,000	1000	1	defibrillatory (1)

4.8 Results of running GSL, AWL, Pilot Science List and new medical lists through the medical corpus

The regularised GSL (word lists GSL1 & GSL2), the AWL, the Pilot Science List, the three MGEN lists, and the twenty-three MED word lists were run through the medical corpus. The lexical profile results of the coverage and occurrence of tokens, types and families across the GSL, AWL, Pilot Science List, MGEN lists, MED lists and words outside the lists are summarised in Table 4.11 (Data in Table 4.11 is based on Appendix 2.1).

The GSL, AWL, and Pilot Science List together provide a coverage of 76.18%, while the three MGEN lists and the twenty-three MED lists cover 21.89% (i.e., 11.96% for the three MGEN lists, and 9.96% for the twenty-three MED lists) of medical texts.

Table 4.11 Cumulative coverage and occurrence of the GSL1, GSL2, AWL, Pilot Science List, the three MGEN lists, and the twenty-three MED lists in the medical corpus

Cumulative results of existing lists	Number of tokens	% of tokens	Number of types	% of types	Number of families
GSL1, GSL2, AWL, Pilot Science List	4137958	76.18	10841	19.59	2944
MGEN (three 1,000) lists	644827	11.87	3000	5.43	0
MED (twenty-three 1,000) lists	540974	9.96	23000	41.63	0
Cumulative total of existing lists	5323759	98.01	36841	66.65	0
Words outside the lists	107981	1.99	18513	33.44	0
Total	5431740	100	55354	100	2944

As the cumulative total Table 4.11 shows, the results of running thirty word lists through the medical corpus using the Range software (Heatley et al., 2002) show that at least 26,000 new medical word types (i.e., three MGEN and 23 MED word lists) need to be added to the GSL, AWL, and Pilot Science List for ESP medical learners to be familiar with 98% (the optimal lexical text coverage figure) of the words they meet when they read medical textbooks in English. (See also Appendix 2 with figures for all 30 lists).

The Range results presented in Table 4.11 seem to make the teaching and learning of the new medical words (at least over 14,000 new medical word types to be taught and learnt assuming our ESP medical students already know all the words included in the GSL, AWL,

and Pilot Science List) a daunting task to undertake given the amount of time, the level of proficiency in English (as we saw in Chapter 3), and the limited content knowledge both ESP students and teachers may have to approach the teaching and learning of thousands of the new medical words.

The expectations of ESP medical undergraduates to learn the large amount of content area (medical) vocabulary needed to get 98% coverage over medical texts seem unrealistic and difficult to achieve in the restricted time span (one or two years at most) of most ESP courses. The vocabulary learning journey across the still unknown words in general academic English before dealing with content specific words that are mostly related to the medical subject domain is an enormous word learning task for ESP medical students.

4.9 The behaviour of the various sets of lists

In order to look at the lexical coverage results of the different word lists (i.e., GSL, AWL, Pilot Science List, and new medical word lists) over the medical and general corpora, we compare (in Tables 4.12 and 4.13) the behaviour of the existing lists. The different behaviour of these lists reflects the different nature of the two corpora.

Table 4.12 Range results for the existing lists (GSL1, GSL2, AWL, and Pilot Science List) for the medical corpus

Word list	Tokens#	Tokens%	Types#	Types%	Families
GSL1	3005823	55.34	4018	7.26	1097
GSL2	352999	6.50	2957	5.34	936
AWL	449900	8.28	2578	4.66	596
Pilot Science List	329230	6.06	1288	2.33	316
Off the lists	1293782	23.82	44513	80.42	0
Total	5431740	100	55354	100	2944

Table 4.13 Range results for the existing lists for the general corpus

Word list	Tokens#	Tokens%	Types#	Types%	Families
GSL1	4181791	76.99	6298	6.45	1132
GSL2	329821	6.07	5297	5.42	1036
AWL	227056	4.18	3307	3.39	600
Pilot Science List	25802	0.48	1231	1.26	314
Off the lists	667270	12.28	81515	83.48	0
Total	5431740	100	97648	100	3082

Next, we compare the Range results for the existing lists (i.e., the GSL1, GSL2, AWL, and Pilot Science List) used in the present chapter over the medical and general corpora.

4.9.1 First thousand GSL coverage over the medical and general corpora

The first 1,132 word families of the GSL (GSL 1) account for 55.34% of the medical corpus and 76.99% of the general corpus. That is, the first thousand word families of the GSL have a striking 21.65% higher coverage for general texts than for medical texts. All the first 1,132 word families of the GSL occur in the general corpus and cover a high percentage of the words in the general corpus (76.99%). Similar results were reported by Coxhead (2000, p. 223) in her AWL study over the academic corpus (71.4%). This shows that the general corpus used as a comparison corpus for this study is quite formal. The general corpus coverage results also show that a corpus consisting of almost five and a half million tokens is easily large enough for every word in the GSL to have a chance of occurring. (See Table 4.14).

Table 4.14 Comparison of range results for the GSL1 for the medical and general corpora

Word list	Corpus	Tokens#	Tokens%	Types#	Types%	Families
GSL 1	medical corpus	3005823	55.34	4018	7.26	1097
	general corpus	4181791	76.99	6298	6.45	1132

A total of 1,097 word families of the first 1,132 word families of the GSL occur in the medical corpus. Some of the headwords occurring with a frequency of 10 or less in the medical corpus were *adventure, beauty, eleven, faith, fellow, Monday, royal, Saturday, story, Thursday, gentleman, yesterday.* Most of the words with low occurrence or zero

frequency of occurrence in the first thousand of the GSL are words needed for telling stories or recounting political and news events.

Thirty-five families in that list did not occur in the medical texts. Examples of the 35 word families of the GSL not occurring (0 frequency) in the medical corpus are: *lady, lord, lovely, Sunday, till, glad, Tuesday, queen, song, dishonour, heaven, joy, captain, empire, fortune, kindly, soul, secretary, victory, literary, landlord, accountant, applicant, noble, poster, waiter, contented, discontent, secrecy, dog, commons, roundabout, disrespect, highness, meaner.*

The first thousand words of the GSL cover only 55.34% of the tokens in the medical corpus. This is very low coverage and is because technical words make up a large proportion of the tokens – around one-quarter of the tokens outside the four traditional word lists are technical terms. This is one word in every four or around 2 to 3 words per line of text. There are medical terms, according to the semantic rating scale developed in this chapter for identifying medical words (see 4.5), in the GSL, AWL, and Pilot Science List (e.g., *blood, body*, *heart*, *sight*) and when these are added to the medical words outside these lists, the coverage of the traditional lists goes down by 14.64% and the coverage of the technical terms gets closer to 36.47% – more than one technical word in every three running words. Because the first thousand GSL included the most frequent words of the English language and accounts for 55.34% of the tokens in medical texts, this is a list of words worth learning for students of medical English. The fact that only 35 word families in the first thousand GSL did not occur at all shows that there is value in learning all the first thousand GSL.

4.9.2 Second thousand GSL coverage over the medical and general corpora

The second thousand GSL (GSL2) accounts for 6.50% of the tokens in the medical corpus and for 6.07% of the general corpus. These figures show roughly similar coverage by the second thousand GSL over both the medical and the general corpora. (See Table 4.15).

Table 4.15 Comparison of range results for the GSL2 for the medical and general corpora

Word list	Corpus	Tokens#	Tokens%	Types#	Types%	Families
GSL2	medical corpus	352999	6.50	2957	5.34	936
	general corpus	329821	6.07	5297	5.42	1036

Out of 1,036 second thousand GSL word families 936 occur in the medical corpus. Examples of the ten most frequent GSL2 word types occurring in the medical corpus are *disease, treatment, risk, patient, severe, during, pain, bone, skin,* and *fever*. Examples of the 267 word families in the second thousand GSL with low frequency (occurring 10 times or less) in the medical corpus are *roar, cape, rival, generous, tribe, ladder, ambition, sincere, conquest, sacrifice, liberty, grammar, verse.*

Examples of the 100 word families in the second thousand GSL not occurring (0 frequency) in the medical corpus are words such as *sorry, ugly, cowardice, delight, thief, brave, cliff, glory, pride, treasure, courage, holy, precious, deed, feast, hatred, voyage, tonight, envious.* Once again, we see that these words are related to narratives and the description of events. These are genres unlikely to occur in medical texts.

The difference in coverage of the second thousand GSL over the medical corpus (6.50%) and the general corpus (6.07%) is just 0.43% higher for medical texts. Given that 100 fewer word families occur in the medical texts, this is a reasonable coverage. Examples of some second thousand GSL words with a medical meaning are *lung, cure, relief, stomach, remedy.* All the second thousand GSL words occurred in the general corpus. The ten most frequent medical words occurring in the second thousand GSL are *disease, treatment, risk, patient, milligram, severe, pain, bone, skin, fever*.

The similarity in coverage of the second thousand GSL for the medical and general corpora indicates that this is a list equally useful for both students of medical English and students of English with general purposes. The coverage results of the second thousand GSL over the medical corpus also suggest that the learning of high-frequency general and academic words in English could be sequenced differently for ESP medical students. Since the coverage by the second thousand words of the GSL is lower than that of the AWL, 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 first thousand GSL.

In general, it may be worth highlighting to medical students which are the medical words in the GSL that occur most frequently in medical texts written in English. Fifty-five GSL first thousand word families occur with a medical meaning. For example, *back, bleed, down, ear, ill, joint, operate, stroke, wound, vessel.* Of the 936 GSL second thousand word

families in the medical corpus, 72 occur with a medical meaning. For example, *health, sick, bone, pain, skin, brain, cough, disease, medicine, patient.*

4.9.3 AWL coverage over the medical and general corpora

The AWL regularised for this study contains 600 word families (compared with 570 in the original list) that account for 8.28% of the 5.4 million running words of the medical corpus. (See Table 4.16). This seems a good coverage of academic words over medicine given the high coverage by technical terms outside the traditional lists. The AWL coverage over medicine (8.28%) is nearly twice as high as the coverage of the AWL over the general comparison corpus (4.18%). The higher coverage by the AWL over the medical corpus shows the academic nature of the medical corpus.

Table 4.16 Comparison of range results for the AWL for the medical and general corpora

Word list	Corpus	Tokens#	Tokens%	Types#	Types%	Families
AWL	medical corpus	449900	8.28	2578	4.66	596
	general corpus	227056	4.18	3307	3.39	600

The coverage by the AWL over medicine (8.63%) is closer to the coverage by the same list over science (9.1%) by Coxhead (2000). Likewise, the AWL coverage over medicine (8.28%) is very similar to the coverage of the second academic corpus (8.5%) also compiled by Coxhead (2000) to serve as a comparison corpus for the AWL study. A coverage of 4.18% by the AWL over the general comparison corpus is reasonably high for academic words in general texts. This moderately high coverage of academic words may have been caused by the inclusion of some formal written documents in various sections of the general corpora and is undoubtedly influenced by the large proportion of newspaper texts, for which the AWL typically provides around 4% coverage (Nation & Webb, 2011, p. 152).

Of the 600 AWL word families, 595 occur in the medical corpus. Of the 595 AWL word families in the medical corpus, 21 occur with a medical meaning: *sex, sufficiency, labour, positive, culture, depress, injure, medical, mental, relax, stress, chemical, media, nuclear, radical, regime, vision, visual, evolution, ratio, abnormal.* The ten most frequent academic words in the medical corpus are *chapter, normal, occur, function, specific, occurs, primary, factors, response, individuals.* Some of the least frequent AWL word families occurring

with a frequency of 10 or less in the medical corpus are *levy, estate, forthcoming, analogy.* The five academic words that do not occur in the medical corpus are *commodity, ideology, subordinate, ministry, ignorant*.

The AWL is a particularly useful word list to learn when ESP students need to focus on academic words. The AWL is also a helpful list for medical students taking first year ESP reading courses.

4.9.4 Pilot Science List coverage over the medical corpus

Of the 317 word families of the Pilot Science List 316 account for 6.06% of the medical corpus and 0.48% of the general corpus of written English. This shows the deliberately narrow range of the words from the Pilot Science List. (See Table 4.17).

Table 4.17 Comparison of range results for the Pilot Science List for the medical and general corpora

Word list	Corpus	Tokens#	Tokens%	Types#	Types%	Families
Pilot Science List	medical corpus	329230	6.06	1288	2.33	316
	general corpus	25802	0.48	1231	1.26	314

Even though nearly all the word families of the Pilot Science List occur in both the medical and general corpora (316 word families occur in the medical corpus, and 314 word families occur in the general corpus); the coverage of this list over the medical corpus is significantly higher (6.06%) than for the general corpus (0.48%). The only word family that did not occur in the medical corpus is *photosynthesis.* Some words in the Pilot Science List, occurring with a frequency of 10 or less in the medical corpus are *dam, drill, extinct, terrestrial.*

This higher coverage of the Pilot Science List over medicine (6.47%) shows that this list plays an important complementary role in helping ESP medical students become familiar with scientific words that occur in texts on health and medicine. Examples of some Pilot Science List words with a medical meaning are *cell, cavity, anatomy,* and *digest.* Likewise, the high coverage shown by the Pilot Science List over the medical corpus suggests that this is a list worth learning for medical students. These results also show that the Pilot

Science List is of particular interest to science and medical students rather than to learners of general English.

Next, we look at the words in the medical corpus not found in any of the existing general or academic word lists, or in the more recently created new medical word lists.

4.10 The words not found in any lists

As shown in the second to last line (in Appendix 2.1), 1.99% of the tokens and 18,513 word types occur in the medical corpus but not in the existing lists or the medical lists. What are these words?

1. Around 20,000 of the 107,981 tokens are single letters of the alphabet or roman numerals.
2. There are also several words that are marginal medical words – *chap* an abbreviation of chapter (3,074 occurrences), *administration* (2,101 occurrences), *prolonged* (1,417 occurrences), and *aggressive* (708 occurrences). Their non-inclusion in the medical word lists is a reflection of this study's policy applied when checking if words were technical or not.
3. Because hyphenated words were separated so that their parts could be treated as simple words, some items not occurring in the lists were prefixes – *anti-*, *non-*, *ex*, *micro-*.
4. The words not in any lists also included some very low-frequency medical terms that also occurred in the general corpus. Examples of these low-frequency medical words are: *neurosis* (1 occurrence in the medical corpus vs. 15 occurrences in the general corpus), *pharmacological* (1 occurrence in the medical corpus vs. 9 occurrences in the general corpus), *encephalographic* (1 occurrence in the medical corpus vs. 1 occurrence in the general corpus), *hematologist* (1 occurrence in the medical corpus vs. 1 occurrence in the general corpus).

Words not in any lists will be looked at in more detail in the following chapter, where the BNC/COCA lists are used to provide a more fine-grained analysis of the frequency levels of the words.

4.11 The medical corpus and general corpus compared

Even though both the medical and general corpora have the same number of running words (5,431,740 tokens for each corpus), the medical corpus has 43% fewer word types (55,354 word types) than the general corpus (97,648 word types). Sutarsyah, Nation and Kennedy (1994) made a similar finding in their comparison of a single economics text (5,438 word families) with a collection of a wide variety of texts (12,744 word families) of similar length. The focus on a single topic area results in a smaller number of different words used – roughly a ratio of 1:2.

A focus on a single topic area also has a striking effect on the frequency of topic-related content words. Table 4.18 shows the top 10 most frequent content words (nouns, verbs, adjectives) in the medical corpus.

Table 4.18 The ten most frequent content words in the medical corpus

Rank	Content words	Frequency in the medical corpus	Rank in the medical corpus	Rank in the general corpus
1	patients	37243	13	1529
2	disease	22662	18	1435
3	treatment	12649	34	1046
4	associated	11116	41	1604
5	therapy	11037	42	8438
6	infection	10914	44	6172
7	cells	10870	45	3869
8	clinical	9671	47	6173
9	risk	9614	48	1392
10	blood	8919	52	866

Table 4.19 The ten most frequent content words in the general corpus

Rank	Content words	Frequency in the general corpus	Rank in the general corpus	Rank in the medical corpus
1	said	12932	47	5601
2	new	9221	63	360
3	time	9159	65	165
4	people	6548	82	1112
5	made	5945	88	540
6	years	5766	90	98
7	man	5664	93	4441
8	way	5290	100	1773
9	good	4693	112	1095
10	work	4600	113	1205

The rank order of the content words in the medical (Table 4.18) and general (Table 4.19) corpora shows that high-frequency content words in the medical corpus are very closely related to topics dealing with health and medicine. By simply looking at this list of ten words in Table 4.18 you could tell what kind of text they came from (e.g., *patients*, *disease*, and *treatment* are words related to the medical field). In the case of the general corpus, there is no particular topic or content area that these top 10 content words in Table 4.19 clearly belong to (e.g., *said, new, time*).

Note also the higher frequency of the top content words in the more focused medical corpus (*patients* 37,243) than in the general corpus (*said* 12,932). The narrow content focus gives these important topic-related words much more opportunity to occur. Note also that the ranking of these content words is much higher in the medical corpus (*patients* – rank13: *said* – rank 47).

With a smaller number of different words in the medical corpus, there is also a smaller number of words occurring only once, twice, and so on (see Table 4.20). Note that in rough approximation to Zipf's law (Sorell, 2012), the percentage figures are somewhat similar.

Table 4.20 Number of word types occurring only once, twice and so on in the medical and general corpora

Occurences	Med1 corpus 5431740 tokens	Med1%	General corpus 5431740 tokens	Gen%
Once	16417	29.66	37812	38.72
Twice	6634	11.98	12500	12.80
3 to 9 times	13586	24.54	23621	24.19
10 to 19 times	5520	9.97	8339	8.54
20 to 49 times	5240	9.47	7325	7.50
50 to 99 times	2822	5.10	3220	3.30
100 to 299 times	2831	5.11	2924	2.99
300 to 499 times	852	1.54	775	0.79
500 to 999 times	725	1.31	602	0.62
1,000 to 9,999 times	682	1.23	471	0.48
10,000 to 99,999 times	40	0.07	53	0.05
100,000 or more	5	0.01	6	0.01
Total	55354	100	97648	100

In relation to the number of word types across frequency bands (occurrences) in the medical and general corpora, the trend in the results indicates that the number of word types for each frequency band is less in the medical corpus than in the general one. The above table also shows that the number of word types with frequencies of one and two in the medical corpus is less than half the number of word types with the same frequencies (one, two) in the general corpus. With word types occurring three and more times the gap between the number of word type occurrences in the medical and general corpora starts to close gradually. Even though the number of word types for each frequency band shown in Table 4.20 is never greater in the medical corpus, from word types occurring 50 or more times in both corpora the frequency gap gets even smaller. When we reach word types with very high frequencies (10,000 times or more) the number of occurrences in both corpora is much closer.

The smaller and more content related proportion of word types found in medical texts underlines the importance of focusing the attention of ESP medical students on the technical vocabulary of their subject area as soon as possible.

4.12 Conclusions

Throughout the present chapter we have looked at the most effective list to move from learning more generally useful vocabulary as represented by the existing word lists, namely,

the GSL, AWL and Pilot Science List to focusing solely on medical vocabulary. This change of focus can be found (1) where readers of medical texts have an adequate general vocabulary size and know the appropriate amount of useful general and academic words (2) where learning the next level of words gives good text coverage, (3) where there is the smallest number of word types to learn to get as close as possible to 98% coverage, and (4) where time is not spent learning general words that do not occur in the medical texts.

The striking features from this study using the existing word lists and lists of medical vocabulary are as follows:

1. There is a very large number of medical words in medical texts – at least 26,000 different word types. This represents an enormous amount of learning for both native speakers and non-native speakers training to be health professionals. These words range from very high-frequency words to many words occurring only once in the corpus.
2. The existing lists perform well on a large medical corpus with most words in the lists occurring in the corpus and with the lists giving reasonable coverage, given the large number of technical terms outside the lists.
3. The contrast in terms of number of different words between a diverse general corpus and a narrowly-focused specialised corpus is striking. The narrowly-focused medical corpus has about half the number of different words, compared to the general corpus. This results in more words with higher frequencies in the specialised corpus in percentage terms. In spite of this, there are still many words to learn.

In the next chapter, we adopt a similar approach to determine the vocabulary load of medical written texts by using another set of more recently developed general word lists (i.e., the BNC/COCA word family lists) to estimate the lexical text coverage of the BNC/COCA word lists over the same medical and general corpora used in the current chapter.

Chapter 5: Medical vocabulary and the BNC/COCA lists

5.1 Introduction and justification

Whereas in Chapter 4 we used the GSL, AWL, and Pilot Science List and created 26 medical word lists to calculate the number of words needed for an optimal coverage (98%) of medical textbooks, in this chapter we use a different set of existing word lists, i.e., the BNC/COCA word family lists (Nation, 2012), to investigate the vocabulary load of the same medical textbooks (i.e., the med1 corpus) examined in Chapter 4 and estimate the number of medical words needed to reach a 98% lexical threshold; the threshold required for good reading comprehension (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt et al., 2011). This chapter has two aims, (1) to see what medical vocabulary is contained in the twenty-five 1,000 word family BNC/COCA lists, and (2) to see what vocabulary size is optimal for students about to undertake medical studies.

Here, the number of medical words across the twenty-five 1,000 BNC/COCA lists is reported. First of all, the lexical coverage results obtained after conducting an independent lexical profile analysis of the same medical and general corpora, of 5,431,740 tokens each, used in Chapter 4 are explained. Most of the lexical profile results discussed in the present chapter have been summarised following Schmitt and Schmitt's (2012) frequency band classification of high, mid and low-frequency words. In the main discussion section of this chapter, the number and percentages of medical words required beyond the limits of each of Schmitt and Schmitt's (2012) frequency bands (high: first 3,000; mid: 4,000-9,000; low:10th 1,000 on) is investigated.

In an effort to determine the number of words needed to reach the 98% lexical threshold of medical texts, several medical word lists are created which go beyond the first three 1,000, the first four 1,000, the first five 1,000, the first six 1,000, the first nine 1,000, the first ten 1,000, and the twenty-five 1,000 BNC/COCA word lists. Furthermore, based on the

average vocabulary size score (6,000 word families) of the 408 participants reported on in Chapter 3, the number of medical words required beyond the first five 1,000 and the first six 1,000 BNC/COCA word lists is also discussed at the end of this chapter. Finally, there are some concluding remarks on the most suitable number of words required to start reading medical texts in English while gaining adequate lexical coverage of medical texts.

5.2 Lexical profiles of the medical and general corpora

In this section, the lexical profiles, in terms of tokens, types and word families of the words occurring in the medical and general corpora compiled for this study are compared. In addition, we also look at the lexical coverage by the proper noun list (Nation, 2005) in these two corpora. In order to investigate the lexical profiles of the medical and general corpora, the Range program (Heatley et al., 2002) has been the software used to run the twenty-five 1,000 BNC/COCA lists first over the medical corpus and then over the general corpus. For more details of lexical profile results from running each of the twenty-five 1,000 BNC/COCA word lists over the medical corpus (Appendix 3.1) and general corpus (Appendix 3.2), see the complete Range outputs in Appendix 3.

5.2.1 Tokens in the BNC/COCA lists

The lexical profile of the occurrence and coverage of the tokens by the BNC/COCA lists over the medical and general corpora organised according to Schmitt and Schmitt's (2012) frequency bands (high, mid and low-frequency word levels), and words outside the lists are summarised in Table 5.1 and Table 5.2.

Table 5.1 The occurrence and coverage of tokens in the medical corpus across the high, mid and low-frequency levels

Frequency bands	Tokens	Cumulative tokens total	% of different tokens by frequency bands	Cumulative % of tokens
High-frequency words 1,000 -3,000	4024861	4024861	74.10	74.10
Mid-frequency words 4,000 -9,000	637796	4662657	11.74	85.84
Low-frequency words 10,000 -25,000	330905	4993562	6.08	91.92
Words outside 25,000	438178	5431740	8.07	100

The lexical text coverage results obtained after running the twenty-five 1,000 BNC/COCA lists over the medical corpus (see Table 5.1) and the general corpus (Table 5.2) separately indicate that at the high-frequency level (1,000 to 3,000) the coverage of the medical texts (74.10%) is much lower (by 14.93%) than the coverage by the same lists of the general texts (89.03%).

At the mid-frequency and low-frequency levels, the lexical coverage of the medical corpus is higher. The twenty-five 1,000 BNC/COCA word lists have a cumulative lexical coverage of 91.92% over the medical corpus and a coverage of 94.67% over the general corpus. Proper nouns which are not technical terms are not included in these figures.

The more detailed results (see Appendix 3.1 and 3.2) show that the biggest difference in token coverage between the two corpora occurs at the first 1,000 level (Medical 51.96%; General 75.59%). This 23.56% lower coverage over the medical corpus is attributed largely to the role played by medical technical words that are not in the first 1,000 BNC/COCA lists or at the lower frequency levels of the BNC/COCA lists. See Tables 5.1 and 5.2.

Table 5.2 The occurrence and coverage of tokens in the general corpus across the high, mid and low-frequency levels

Frequency bands	Tokens	Cumulative tokens total	% of different tokens by frequency bands	Cumulative % of tokens
High-frequency words 1,000 -3,000	4836071	4836071	89.03	89.03
Mid-frequency words 4,000 -9,000	258544	5094615	4.75	93.78
Low-frequency words 10,000 -25,000	47807	5142422	0.88	94.66
Words outside 25,000	289318	5431740	5.33	100

From the second 1,000 to the twenty-fifth 1,000 BNC/COCA list, the text coverage of each subsequent BNC/COCA list is always higher over the medical corpus than over the general corpus (see Appendix 3.1 and 3.2). The explanation for the higher coverage of the medical corpus by 24 of the 25 BNC/COCA lists is that medical technical terms occur frequently in all these lists (e.g., the coverage of the BNC 2nd 1,000 is 11.68% over the medical corpus vs. 8.55% over the general corpus; the coverage of the BNC 3rd 1,000 is 10.46% over the medical corpus vs. 4.89% over the general corpus; the coverage of the BNC 4th 1,000 is 4.16% over the medical corpus vs. 1.84% over the general corpus, and so forth). In Chapter 4 which looked at the coverage by the GSL and AWL, we also found that the biggest difference in token coverage between the medical and general corpora occurs at the GSL1 frequency level. Likewise, the text coverage results of each subsequent word list in Chapter 4, namely, the GSL2, AWL and Science list, are always higher over the medical corpus. For example, the coverage by GSL2 is 0.43% higher, the coverage by the AWL is 4.10% higher, and the coverage by the Science list is 5.58% higher over the medical corpus than over the general corpus.

5.2.2 Word types in the BNC/COCA lists

Table 5.3 summarises the number of word types for each corpus and the percentage of occurrences of the word types for the high, mid, and low-frequency bands in the twenty-five 1,000 BNC/COCA lists.

Table 5.3 The occurrence of word types and families in the medical corpus (Med1) and general corpus (Gen) across the high, mid and low-frequency levels

Frequency bands		Word types	Cumulative word types total	% of word types by frequency levels	Word families	Medical and general examples
High-frequency words 1,000 -3,000	Med1	10345	10345	18.69	2847	life, depression, joint
	Gen	16340	16340	16.74	2999	amount, basis, determine
Mid-frequency words 4,000 -9,000	Med1	7897	18242	14.26	3839	diagnosis, obesity, invasive
	Gen	17690	34030	18.12	8977	classify, disadvantage, factual
Low-frequency words 10,000 -25,000	Med1	5625	23867	10.16	4126	hepatic, cutaneous, fibromas
	Gen	12314	46344	12.60	17526	brevity, transitory, confluence
Words outside 25,000	Med1	31487	55354	56.88	0	pathologic, anatomic, pericarditis
	Gen	51304	97648	52.54	0	cultivable, pertinence, investigational

There is a much smaller number of word types in the medical corpus (55,354 word types) than in the more diverse general corpus (97,648 word types). Throughout the twenty-five 1,000 BNC/COCA lists, the Range results of the medical and general texts consistently show a smaller number of word types occurring in the medical corpus than in the general one in each of the twenty-five 1,000 BNC/COCA lists. These figures clearly illustrate the restricted nature of the vocabulary found in the medical corpus and highlight the need to focus the attention of native and non-native readers of medical texts on this more limited set of specialised words. Essentially, the medical corpus uses a much smaller vocabulary at all levels, but uses many of these different technical words frequently.

5.2.3 Word families in the BNC/COCA lists

As also shown in Table 5.3, this more restricted number of word types encountered in the medical corpus parallels the number of word families. There is a smaller number of word families from the twenty-five 1,000 BNC/COCA lists occurring in the medical corpus compared with the general corpus. For example, while 999 BNC/COCA word families from the same first 1,000 BNC/COCA list occur in the general corpus, 973 word families from the first 1,000 BNC/COCA word list occur in the medical corpus. Another example is at the seventeenth 1,000 BNC/COCA frequency level where only 261 medical corpus word families occur compared to more than half the general corpus word families (552) occurring at the same frequency level. The narrowly focused content of the medical corpus means that fewer words occur at each of the twenty-five 1,000 word levels in the BNC/COCA lists. Similarly, with regard to the total number of word families from the twenty-five BNC/COCA lists, there are 17,512 in the general corpus, and 10,811 word families in the medical corpus, a difference of 6,701 word families.

5.2.4 Lexical coverage by the proper noun list in the medical and general corpora

The results of adding the proper noun list (Nation, 2005) to the twenty-five 1,000 BNC/COCA word family lists when looking at the lexical profile of these lists over the medical and general corpora show that the proper noun list covers 3.33% of the tokens in the general texts, and 0.98% in the medical. This 0.98% coverage result over medical texts is, however, a disputable figure considering some of the proper nouns in the medical corpus have a specific medical meaning that differs from their meaning in general English (e.g., Hodgkin, Alzheimer in Table 5.4 below).

Table 5.4 The coverage, and occurrence of the BNC/COCA proper noun list over the medical and general corpora

Proper nouns	Tokens%	Number of tokens	Number of types	Number of families	Examples
Medical corpus	0.98	53457	1885	1774	Hodgkin, Alzheimer, HIV
General corpus	3.33	180997	13361	12932	York, Americas, Murphy

The examples in Table 5.4 of frequent proper nouns in the corpora show that the proper nouns found in medical texts are frequently used to refer to diseases and medical conditions. However, proper nouns in the general corpus are used to name people and places. This distinction in the use and meaning of proper nouns in the medical and general texts makes it worth considering the possibility of making a list of proper nouns frequently used in specialised medical texts.

In the next section, the frequency of occurrence of the medical words across the BNC/COCA lists is discussed.

5.3 Medical words across the twenty-five 1,000 BNC/COCA lists

In order to find out where the medical words occur across the BNC/COCA lists, using the Range software (Heatley et al., 2002), the twenty-five 1,000 BNC/COCA lists were run over the list of the 32,194 most frequent word types classified as medical words (see 4.5) in the medical corpus. Table 5.5 summarises the total lexical coverage and the cumulative coverage across Schmitt and Schmitt's (2012) frequency band classification. Some examples of medical words belonging to these frequency bands are also provided.

Table 5.5 The percentage occurrence of medical technical word types across the high, mid and low-frequency levels

Frequency bands	% of different word types by frequency levels	Cumulative %	Examples
High-frequency words 1,000 - 3,000	8.67	8.67	blood, cure, cancer
Mid-frequency words 4,000 - 9,000	10.47	19.14	tumor, anemia, catheter
Low-frequency words 10,000 - 25,000	14.81	33.95	amnesia, acidophilus, gonococcal
Medical words outside 25,000	66.07	100.02	Cor, ebola, venulitis

Note in Table 5.5 that the majority of the word types (66.07%) occur outside the twenty-five 1,000 levels, indicating how specialised these words are, and how unlikely they are to be known by non-medical people.

Just under one technical medical word in every ten medical word types is a high-frequency word (8.67%) and just under one in every five medical word types (19.14%) is from the high or mid-frequency word lists. Over 80% are low-frequency words (80.88%). This figure agrees with Chung and Nation's (2004) finding for the sources of vocabulary in an anatomy text.

The raw number of medical word types and medical word families occurring across the three frequency bands is shown in Table 5.6. These, of course, correspond to the percentages in Table 5.5.

Table 5.6 Distribution of the medical technical word types (raw figures) across the major frequency levels

Frequency bands	Word types	Cumulative total of word types	Word families total	Cumulative total of word families
High-frequency words 1,000 - 3,000	2792	2792	554	554
Mid-frequency words 4,000 - 9,000	3370	6162	753	2098
Low-frequency words 10,000 - 25,000	4762	10924	3394	5492
Medical words in the proper nouns, marginal words, transparent compounds and abbreviations BNC/COCA lists	21271	32194	1484	6976

A total of 10,924 medical word types, and 5,492 medical word families occur in the twenty-five 1,000 BNC/COCA lists. Over 21,271 medical word types, and over 1,484 word families are outside the twenty-five 1,000 BNC/COCA word lists (i.e., the proper nouns, marginal words, transparent compounds and abbreviations lists), indicating that anyone studying medicine has a large amount of vocabulary to learn that is unlikely to be known from their general vocabulary knowledge of English. Figures for each of the twenty-five 1,000 BNC/COCA lists can be found in Appendix 3.5. The 21,271 medical word types

outside the BNC/COCA lists have not been grouped into word families, because, as we have already discussed in Chapter 4, the word type is the unit of counting selected in the present study to create new medical word lists, and estimate the vocabulary load of medical textbooks.

According to the results of the VST administration to adult native speakers of English, and to EFL learners, the average vocabulary size of these two groups is 20,000 word families (Nation, 2006) for native speakers of English and 6,000 word families (Elgort, 2013; Karami, 2012; Nguyen & Nation, 2011) for non-native speakers of English. These vocabulary size estimates suggest that first year medical students who are native speakers of English with an average vocabulary size of 20,000 word families still need to learn over 21,271 medical word types outside the twenty-five 1,000 BNC/COCA lists. Meanwhile, EFL learners taking their first English for Medical Purposes course at university have a more difficult task. They need to learn many of the mid-frequency and all of the low-frequency medical words as well.

5.3.1 Distribution of medical words across the twenty-five 1,000 BNC/COCA lists

Table 5.7 summarises the number and percentage of BNC/COCA medical word families occurring in the medical and general corpora used for the present study. The results have also been grouped using Schmitt and Schmitt's (2012) frequency band classification. Some examples of medical words belonging to these frequency bands are also provided.

Table 5.7 Total word families appearing in the medical and general corpora

BNC word family lists	Medical word families	% of medical word families in the bands	Total medical and general word families	Examples
High-frequency words 1,000 -3,000	502	16.70	3000	cell, brain, gene
Mid-frequency words 4,000 -9,000	1204	20.10	5997	insulin, serum, tumor
Low-frequency words 10,000 -25,000	3053	28.70	10641	dialysis, herpes, venous
Total	4759	24.20	19638	

Table 5.7 shows that 502 of the first 3,000 word families (16.7%) are medical word families, and that there are 1,204 medical word families (20.1%) in the mid-frequency band, and 3,053 medical word families (28.7%) in the low-frequency band. The highest percentage (64.2%) of medical word families (3,053) is in the low-frequency band while there is a total of 1,706 (35.8%) medical word families in the high and mid-frequency bands. (See Appendix 3.6 with number and percentage of medical word families for each of the twenty-five 1,000 BNC/COCA word family lists).

As shown in Table 5.7, of the total number of word families occurring in the medical and general corpora (19,638 word families), 45.8% (8,997 word families) of these 19,638 BNC/COCA word families occur in the high and mid-frequency bands while 54.2% (10,641 word families) occur in the low-frequency band. 5,362 (21.4%) of the 25,000 BNC/COCA word families do not occur in any of the two corpora compiled for the present study. These are all in the low-frequency band.

Table 5.8 summarises the word type results of the BNC/COCA lists in the medical corpus. As we can see in column 2 of Table 5.8, there are 10,931 medical word types in the twenty-five 1,000 BNC/COCA lists which occur in the medical corpus. This corresponds to 14.4% of the total number of word types (75,679) in the twenty-five 1,000 BNC/COCA lists.

Table 5.8 Number and percentage of medical word types in the medical and general corpora across the twenty-five 1,000 BNC/COCA lists

Frequency bands	Medical word types by frequency bands	% of medical word types in BNC/COCA	Total word types in BNC/COCA lists	Word types occurring in the medical and general corpora
High-frequency words 1,000 -3,000	2792	14.60	19107	16682
Mid-frequency words 4,000 -9,000	3377	14.30	23555	18474
Low-frequency words 10,000 -25,000	4762	14.40	33017	15752
Total	10931	14.40	75679	50908

Of the 75,679 word types (see Table 5.8, column 4) in the twenty-five BNC/COCA lists, 67.3% (50,908) occur in the medical and general corpora, and 32.7% (24,771) do not occur in any of the two corpora (medical and general).

Furthermore, 79,499 word types in the medical and general corpora do not occur in any of the twenty-five 1,000 BNC/COCA lists. These words consist of 22,985 classified word types (1,715 classified as general word types and 21,270 classified as medical word types), and 56,514 unclassified word types with frequency values in the medical corpus of one or zero. This study left 56,514 word types unclassified because of their low frequency in the medical corpus (see Appendix 3.7 for more details).

5.3.2 Medical words in the medical corpus

In this section, the amount of the medical words (tokens, word types, and word families) across the twenty-five 1,000 BNC/COCA lists is discussed. The results for this part of the discussion are also presented using Schmitt and Schmitt's (2012) frequency band classification. Table 5.1 looked at all the tokens occurring in the various bands regardless of whether they were medical words or not. Table 5.9 (columns 2 and 3) considers only the tokens that were medical words.

Table 5.9 Number and percentage of medical tokens across the twenty-five 1,000 BNC/COCA lists in the medical corpus

BNC/COCA word-family lists by frequency bands	Medical tokens by frequency bands	Cumulative total of medical tokens	Percentage of medical tokens	Cumulative percentage of medical tokens
High-frequency words 1,000 -3,000	709265	709265	13.06	13.06
Mid-frequency words 4,000 -9,000	558083	1267348	10.26	23.32
Low-frequency words 10,000 -25,000	325677	1593025	5.99	29.31

Table 5.9 shows that within the BNC/COCA lists the highest number of medical tokens (709,263) and the highest lexical text coverage (13.06%) occurs with words from the high-frequency band, and the lowest number of medical tokens (325,677) and lowest lexical coverage (5.99%) occurs with words from the low-frequency band.

There are 1,593,025 medical tokens in the medical corpus from the twenty five BNC/COCA lists, and these account for 29.31% of the tokens used in the medical corpus.

As summarised in Table 5.10 below, the number of medical word types from the twenty-five 1,000 BNC/COCA lists is as follows: 2,076 medical word types at the high-frequency level, 3,055 medical word types at the mid-frequency level and 4,610 medical word types at the low-frequency level. The cumulative total of medical word types across the high, mid, and low-frequency bands is 9,741. In addition, there are at least 21,250 medical word types outside the BNC/COCA lists.

Table 5.10 Number of word types and word families across the twenty-five 1,000 BNC/COCA lists

Frequency bands	Medical types	Cumulative total of medical types	Medical families	Cumulative total of medical families	Examples
High-frequency words 1,000 -3,000	2076	2076	495	495	bone, muscle, cancer
Mid-frequency words 4,000 -9,000	3055	5131	1137	1632	liver, plasma, fibrosis
Low-frequency words 10,000 -25,000	4610	9741	2991	4623	hyperglycemia, pituitary, angina, ocular
Words outside 25,000	21250	30991	0	0	antimicrobial, pathologic, intravascular

Let us now look at the medical word families in the twenty-five 1,000 BNC/COCA lists (Table 5.10): 495 word families occur at the high-frequency level, 1,137 families at the mid-frequency level, and 2,991 families at the low-frequency level. Since no word families have been created outside the twenty-five 1,000 BNC/COCA lists for the present study, it is not possible to calculate the number of word families outside the BNC/COCA lists.

There is a total of 4,623 medical word families in the twenty-five 1,000 BNC/COCA lists. These 4,623 word families correspond to just under one-fifth (18.49%) of the 25,000 word families in the BNC/COCA lists. Medical words clearly play an important role in everyday vocabulary use and thus students of medicine will come to their study being partly familiar with the vocabulary of that specialist area. However, these more commonly known words make up only a small part of a medical vocabulary.

Let us now look at the relationship between the vocabulary size of prospective medical students and the number of words needed for achieving the 98% lexical threshold required for reading medical texts will be described.

5.4 The vocabulary size needed to begin medical study

We have now looked at the occurrence of medical words in the BNC/COCA lists. We can use this information to look at the vocabulary load of medical texts for learners with various vocabulary sizes. This will help us address the question of the minimum number of words a learner needs to know before embarking on medical study.

The procedure used to see how many words are needed involves choosing a particular vocabulary size, say for example 3,000 word families, and then seeing how many words would have to be learned to gain 98% coverage of medical text. By comparing various vocabulary sizes, we can see which size provides the best balance between learning a mixture of general and medical vocabulary and concentrating only on medical vocabulary. Is this move to a focus on only medical vocabulary best done when learners know 3,000 words, 4,000 words, and 5,000 words and so on? Answering this question involves making medical word lists that begin at each of the vocabulary size levels investigated. So, for a vocabulary size of 3,000 word families for example, it is necessary to exclude the medical words occurring in the most frequent 3,000 word families (because they are assumed to be known) and create medical lists that begin from that level.

5.5 Medical words beyond the first three 1,000 BNC/COCA word lists

Twenty-seven medical word lists first created in Chapter 4 were used in this part of the study. They included the four 1,000 medical and general (MGEN) word lists obtained from a list of 4,149 medical word types with frequency values in the medical corpus equal to or greater than six and up to 9,370, and twenty-three 1,000 medical only (MED) word lists from a list of 23,489 medical word types with frequency values in the medical corpus ranging from 3,194 to 1 (see section 4.7).

Table 5.11 shows the number and percentage of the word tokens in the three 1,000 most frequent BNC/COCA word lists, the four 1,000 MGEN lists, and the twenty-three 1,000 MED lists.

Table 5.11 Range results on the medical corpus using the first three 1,000 BNC/COCA lists

WORD LIST	TOKENS#	TOKENS%	TYPES#	TYPES %	FAMILIES
BNC 1st 1,000	2822197	51.96	3648	6.59	973
BNC 2nd 1,000	634329	11.68	3374	6.10	941
BNC 3rd 1,000	568335	10.46	3323	6.00	933
MGEN 1st 1,000	492781	9.07	1000	1.81	1000
MED 1st 1,000	286150	5.27	1000	1.81	1000
MGEN 2nd 1,000	129322	2.38	1000	1.81	1000
MED 2nd 1,000	81207	1.50	1000	1.81	1000
MGEN 3rd 1,000	63382	1.17	1000	1.81	1000
MED 3rd 1,000	45478	0.84	1000	1.81	1000
MGEN 4th 1,000	36561	0.67	1000	1.81	1000
MED 4th 1,000	29880	0.55	1000	1.81	1000
MED 5th 1,000	21478	0.40	1000	1.81	1000
MED 6th 1,000	16411	0.30	1000	1.81	1000
MED 7th 1,000	12807	0.24	1000	1.81	1000
MED 8th 1,000	10172	0.19	1000	1.81	1000
MED 9th 1,000	8330	0.15	1000	1.81	1000
MED 10th 1,000	6716	0.12	1000	1.81	1000
MED 11th 1,000	5633	0.10	1000	1.81	1000
MED 12th 1,000	4868	0.09	1000	1.81	1000
MED 13th 1,000	4000	0.07	1000	1.81	1000
MED 14th 1,000	3599	0.07	1000	1.81	1000
MED 15th 1,000	3000	0.06	1000	1.81	1000
MED 16th 1,000	3000	0.06	1000	1.81	1000
MED 17th 1,000	2085	0.04	1000	1.81	1000
MED 18th 1,000	2000	0.04	1000	1.81	1000
MED 19th 1,000	2000	0.04	1000	1.81	1000
MED 20th 1,000	2000	0.04	1000	1.81	1000
MED 21st 1,000	1453	0.03	1000	1.81	1000
MED 22nd1,000	1000	0.02	1000	1.81	1000
MED 23rd 1,000	1000	0.02	1000	1.81	1000
Off BNC MGEN MED lists	130566	2.40	18009	32.53	0
Total	5431740		55354		29847

Note in Table 5.11 that these lists reach 97.6% (close to 98%) token coverage of the medical corpus.

The most efficient point to move from learning generally useful vocabulary as represented by the BNC/COCA lists to focusing solely on medical vocabulary is (1) where readers of medical texts have an adequate general and academic vocabulary size, (2) where learning the next level of words gives good text coverage, (3) where there is the smallest number of word types to learn to get as close as possible to 98% coverage, and (4) where time is not spent learning general words that do not occur in the medical corpus. Let us look at each of these criteria in turn.

In Table 5.12 we can see that the first 3,000 word families of the BNC/COCA cover 74.10% of the tokens. If the learner then moves to learning medical words instead of going on to the 4th 1,000 of the BNC/COCA lists, the MGEN 1st 1,000 list gives an additional 9.07% coverage.

Table 5.12 Number of word types needed to get close to 98% coverage

Assumed vocabulary size (BNC/COCA word family lists used)	Number of word types in the assumed size	Percentage of token coverage by the assumed vocabulary size	Number of 1,000 medical word type lists outside the BNC/COCA lists	Number of word types outside the assumed size	Total number of word types	Total percentage of tokens
3,000	10345	74.10	27	27000	37345	97.63
4,000	12440	78.26	16	16000	28440	97.91
5,000	14110	80.83	12	12000	26110	97.94
6,000	15503	82.54	11	11000	26503	98.06
9,000	18242	85.84	9	9000	27228	98.04
10,000	18841	86.57	9	9000	27841	98.09
25,000	23867	91.92	6	6000	29853	98.18

From the point of view of number of types to learn, knowing 5,000 word families is the best size, because as Table 5.12 shows there would be an additional 12,000 medical word types to learn (beyond the 14,110 word types in the first five 1,000 BNC/COCA lists) compared to 16,000 medical word types if 4,000 word families (12,440 word types) were known, and 11,000 medical word types if 6,000 word families (15,503 word types) were known. Although knowing 6,000 word families plus the medical words gives slightly higher text coverage (98.06% compared with 97.94%), this extra text coverage is not as valuable as the time saving by not learning the 6th 1,000 words.

As Table 5.13 indicates (see also Appendix 3.9), learning the 4th 1,000 of the BNC/COCA would only give an additional 4.16% coverage, but 839 word families of the 4th 1,000 list do occur in the medical corpus. After the first 3,000 of the BNC/COCA which comprises 10,345 word types, an additional 27,000 word types are needed to get 97.63% coverage.

Table 5.13 Coverage provided by the next list of words

Assumed vocabulary size (BNC/COCA word family lists used)	Percentage % coverage by the next 1000 BNC/COCA words	Coverage by the next 1000 medical words
3,000	4.16 (4th 1000)	9.07**
4,000	2.57 (5th 1000)	7.17
5,000	1.71 (6th 1000)	5.82
6,000	1.57 (7th 1000)	5.15
9,000	0.73(10th 1000)	4.86
10,000	0.67 (11th 1000)	4.73
25,000	Less than 0.06	2.59

With regard to coverage by the next list of words and number of words occurring in the next list of words, the earlier medical word learning occurs, the better it is for the learner. This is because the early lists of the medical words are those that occur very frequently and of course all the words in the medical lists occur in the text they were made from. Note in Table 5.14 column four below that roughly one-quarter of the words in the next BNC/COCA list are medical words.

The disadvantage of focusing on medical words too early is that the general words that occur in the medical text as well as academic words which are not medical, may not be learnt and these words do occur often in medical texts. As Table 5.14 suggests (column 2 minus column 4), up to the 6,000 word level, several hundred words from the BNC/COCA lists occurring in the medical corpus.

Table 5.14 Number of words occurring in the next 1000 words

Assumed vocabulary size (BNC/COCA word family lists used)	Number of word families occurring in the next 1000 BNC/COCA words	Number of word types occurring in the next 1000 medical words	Number of medical word families in the next 1,000 BNC/COCA words (from Appendix 1.5)
3,000	839 (4th 1000)	1000	254 (4th 1000)
4,000	764 (5th 1000)	1000	283 (5th 1000)
5,000	677 (6th 1000)	1000	254 (6th 1000)
6,000	593 (7th 1000)	1000	249 (7th 1000)
9,000	396 (10th 1000)	1000	233 (10th 1000)
10,000	384 (11th 1000)	1000	241 (11th 1000)
25,000	Less than 384	1000	

Tables 5.12, 5.13, and 5.14 are based on data from Appendixes 3.8 to 3.14.

So, considering the number of word types to learn and the need to keep learning general words, it would be best if learners had a vocabulary size of at least 5,000 or 6,000 word families before beginning medical study. Even with this vocabulary size, there are still thousands of medical words to learn.

5.6 Conclusions

The data gathered in this chapter supports the findings of Chapter 4, namely, that technical words make up a very large proportion of the tokens, types and word families in the medical textbooks, but that the number of different words occurring in the medical corpus is substantially less than the number of different words in the similarly sized general corpus.

This chapter shows that there is a spread of medical words through all the twenty-five 1,000 BNC/COCA word family lists. It also shows that, even with a very large general vocabulary (as represented by the BNC/COCA lists), there would still be a staggering number of medical words to learn to reach 98% lexical coverage. The amount of these medical words is so large that learners studying medicine need to get on to learning them as soon as possible. It seems that a vocabulary size of at least 5,000 word families is necessary for second language learners of English.

Chapter 6: The validation of the medical word lists on an independent medical corpus (the med2 corpus)

6.1 Introduction and justification

This aim of this chapter is to report on the lexical profile results from running the same existing word lists (i.e., the GSL, AWL, Pilot Science List, and BNC/COCA lists), and the medical word lists created in Chapters 4 and 5 on a new independent medical corpus (the med2 corpus). The main purpose for running these existing general, academic and specialised medical word lists on a new medical corpus is to investigate their behaviour in a different medical corpus from the one used to make them and thus check the reliability of the lists.

In the first part of this chapter, the compilation process of the med2 corpus is explained. Then, comparisons of the lexical text coverage results of the med1 and med2 corpora are made. These coverage comparisons are made first using the GSL, AWL, and Pilot Science List, and then using the twenty-five 1,000 BNC/COCA lists. The discussion of the similarities and differences of the lexical profiles of the med1 and med2 corpora continues with an analysis of the coverage results of the medical lists created in Chapters 4 and 5. The chapter concludes by comparing the distribution of frequencies across the med1 and med2 corpora.

6.2 Compilation of a new medical corpus

The main features of a new independent medical corpus (the med2 corpus) compiled for validating the medical word lists developed in Chapters 4 and 5 are summarised in Table 6.1.

Table 6.1 Features of the two medical texts used for making the med2 corpus

Features	Merck textbook	Oxford textbook
Length	1905305	3985172
Number of chapters/sections	353 chapters written by different authors	33 sections written by different authors. Each section includes several subsections.
English variety	American English	British English
Genre	Written academic English	Written academic English

The compilation procedure and the inclusion and exclusion criteria followed to create this new independent medical corpus (with a total of 5890477 tokens) are described in Table 6.2. The specialised texts included in the two medical textbooks used to create the med2 corpus for the present study were downloaded from various publishers and databases such as Merck Sharp and Dohme Corp, Academic OneFile, and ProQuest, made available by Victoria University of Wellington for research purposes.

Table 6.2 Compilation of the two medical textbooks (the med2 corpus)

Compilation	Merck textbook	Oxford textbook
Compilation procedure	Convert the electronic version (chm files) to plain text, and delete the information listed in the exclusion criteria.	Convert the digital version (chm files) to plain text, and delete the information listed in the exclusion criteria.
Inclusion criteria	353 chapters, 3 appendixes, 1 index	33 main sections with several subsections, 1 index
Exclusion criteria	Cover, title page, preface, copyright information, list of contributors, acknowledgements, all bibliographical references, names of authors, URL addresses.	Cover, title page, front matter, copyright information, foreword, acknowledgements, bibliographical references, names of authors, URL addresses.

As Table 6.2 shows, the med2 corpus is the result of combining the electronic versions of two textbooks on the fundamentals of medicine, namely, the *Merck Manual of Diagnosis and Therapy* , 19th edition (Porter & Kaplan, 2011) and the *Oxford Textbook of Medicine*, 5th edition (Warrell, Cox, Firth, & Ogg, 2010). These two textbooks were chosen because they cover a wide range of medical topics, and cover similar topics to the ones included in the med1 corpus (the medical corpus used to create the medical word lists that will be evaluated in this chapter). Another reason for choosing these two medical textbooks for the

new independent medical corpus is that they are books commonly consulted by medical students from the beginning of their medical studies. Overall, the med2 corpus contains 458,737 more tokens than the med1 corpus. The larger size of the med2 corpus is due to the fact that the full texts with all the chapters were used when compiling both the med1 and med2 corpora.

6.3 Lexical text coverage of the med2 corpus

The lexical profile of the med2 corpus was found by running several existing word lists over the new 5.8 million medical corpus using the Range software (Heatley et al., 2002). Firstly, the lexical text coverage of the GSL1, GSL2, AWL and Pilot Science List on the med2 corpus is presented. Then, the coverage results using the twenty five 1,000 BNC/COCA lists on the same medical corpus (the med2 corpus) are discussed.

6.3.1 Lexical text coverage of the med2 corpus using the GSL1, GSL2, AWL, and Pilot Science List

The lexical profiles of the med1 and med2 corpora are both included in the same table (Table 6.3) to allow easy comparison of the lexical coverage of both corpora by the GSL, AWL, and Pilot Science List.

Table 6.3 The occurrence and coverage of tokens, word types and word families in the med1 and med2 corpora by the GSL, AWL, and Pilot Science List

Word lists		Number of tokens	% of tokens	Number of types	% of types	Number of families	Examples
GSL1	Med1	3005823	55.34	4018	7.26	1097	bleeding, strokes
	Med2	3280826	55.70	4384	6.75	1110	
GSL2	Med1	352999	6.50	2957	5.34	936	fever, stomach
	Med2	409297	6.95	3365	5.18	982	
AWL	Med1	449900	8.28	2578	4.66	596	depression, injure
	Med2	454065	7.71	2709	4.17	599	
PSL	Med1	329230	6.06	1288	2.33	316	plasma, tract
	Med2	347128	5.89	1332	2.05	316	
Off the lists	Med1	1293782	23.82	44513	80.42	0	coronary, marrow
	Med2	1399161	23.75	53129	81.84	0	
Total	Med1	5431740	100	55354	100	2944	
	Med2	5890477	100	64919	100	3007	

Table 6.3 shows that the lexical profiles of the med1 and med2 corpora are similar across the four word lists. To begin with, the lexical text coverage of the GSL1 and GSL2 on the

med2 corpus is barely 0.5% higher than the coverage of the med1 corpus by the same word lists. With respect to the coverage of the AWL and Pilot Science List over both medical corpora (i.e., the med1 and med2 corpora), they are respectively only 0.57% and 0.17% lower over the med2 corpus. Even the difference in coverage of the tokens found in the words outside these four word lists is insignificantly small, with just 0.07% lower coverage on the med2 corpus.

The fact that the med2 corpus has 458,737 more tokens than the med1 corpus makes the word type results of the med2 corpus consistently higher across the GSL, AWL and Pilot Science List. When we compare the total number of word types in both corpora the med2 corpus has 9,565 (15%) more word types than the med1 corpus.

In relation to the number of word families in the med1 and med2 corpora, there are only 63 (3%) more word families in the med2 corpus distributed as follows: 13 (2%) more families in GSL1, 46 (5%) more in GSL2, and 3 (1%) more in AWL, and the same number of word families (316) in the Pilot Science List. In other words, between 96% and 98% of the total 1,132 GSL1 word families, 90% and 95% of the 1,036 GSL2 word families, 99% of the 600 AWL word families, and 99.7% of the 317 Pilot Science List word families occur in the med1 and med2 corpora, respectively.

6.3.2 Lexical text coverage of the med2 corpus by the twenty five 1,000 BNC/COCA lists

The lexical text coverage results of the med2 corpus (compared with the med1 corpus) by the twenty five 1,000 BNC/COCA lists are summarised in Table 6.4.

Table 6.4 The occurrence and coverage of tokens, types and families in the med2 corpus and (the med1 corpus) across the high, mid and low-frequency levels

Frequency bands		Number of tokens	% of tokens	Number of types	% of types	Number of families	Examples
High- frequency words 1,000 -3,000	Med1	4024861	74.10	10345	18.69	2847	ageing, brain
	Med2	4362963	74.07	11386	17.53	2922	
Mid-frequency words 4,000 -9,000	Med1	637796	11.74	7897	14.26	3839	fetus, colitis
	Med2	689233	11.69	9083	13.99	4305	
Low-frequency words 10,000 -25,000	Med1	330905	6.08	5625	10.16	4126	aciduria, herpes
	Med2	360330	6.14	6529	10.06	4736	
Cumulative total of frequency bands	Med1	4993562	91.92	23867	43.11	10812	
	Med2	5412526	91.90	26998	41.58	11963	
Words outside 25,000	Med1	438178	8.07	31487	56.88	0	cytokines, eclampsia
	Med2	477951	8.11	37921	58.41	0	
Total	Med1	5431740	100	55354	100	10812	
	Med2	5890477	100	64919	100	11963	

In spite of the med2 corpus having 458,737 more tokens than the med1 corpus, the twenty-five 1000 lists cover 91.9% of both medical corpora, and thus for words beyond the twenty five 1,000 BNC/COCA lists, the lexical text coverage of both medical corpora is around 8.1%. A similar trend is observed in the fourth column of Table 6.4 giving the number of word types, where even the slightly higher number of the med2 corpus word types provides a similar percentage of types across the high, mid, and low-frequency bands. With regard to word families, the numbers are rather similar between the two corpora given that the med2 corpus is a longer corpus.

6.3.3 Lexical text coverage by BNC/COCA proper noun list on the medical corpora

Let us now look at the lexical text coverage results of the BNC/COCA proper noun list over the med1 and med2 corpora. As Table 6.5 shows, the lexical text coverage of the proper noun list over the med1 and med2 corpora is very similar.

Table 6.5 The coverage, and occurrence of the BNC/COCA proper noun list over the med1 and med2 corpora

Proper nouns	Tokens%	Number of tokens	Number of types	Number of families	Examples
Medical corpus 1	0.98	53457	1885	1774	Doppler, HLA, Pacific, California
Medical corpus 2	0.96	5427	2452	2303	Parkinson, ANA, British, Caribbean

In spite of the med2 corpus having 2,970 more tokens, 567 more word types, and 529 more word families than the med1 corpus, the BNC/COCA proper noun list covers nearly 1% (0.98% of the med1 corpus and 0.96% of the med2 corpus) of the medical texts included in the two corpora compared in Table 6.5.

The examples in column six of Table 6.5 illustrate the inclusion of proper nouns with either a general or a more medical meaning in the proper noun BNC/COCA list. For example, nouns like Doppler, Parkinson, and acronyms like HLA (human leukocyte antigen) and ANA (an antinuclear antibody) have a specialised use in the medical texts while proper nouns like Pacific, California, British and Caribbean have a more general use.

Here, we have compared the lexical text coverage results of the existing general/academic lists (the GSL, AWL, Pilot Science List, and BNC/COCA lists), and the new medical word lists created in Chapters 4 and 5 over both the med1 and med2 corpora, finding very similar coverage. In the next section (6.4) we examine some criteria to decide on the best lexical moment to move from learning general vocabulary to acquiring a larger repertoire of academic, scientific and more specialised words also required to reach a good lexical threshold of medical texts written in English.

6.4 Reasons for embarking on the learning of medical vocabulary

It is important to consider once again some of the reasons why readers of medical texts need to move from learning generally useful vocabulary to focusing solely on learning medical vocabulary. As has already been stated in previous chapters, this research takes

into account four criteria to decide on the most efficient point to move from learning generally useful vocabulary, as represented by the GSL, AWL, and Pilot Science List (in Chapter 4) and the BNC/COCA lists (in Chapter 5), to focusing only on medical vocabulary. These four criteria include the point (1) where readers of medical texts have an adequate general vocabulary size and know a useful amount of the general and academic words, (2) where learning the next level of words gives the best text coverage, (3) where there is the smallest number of word types to learn to get as close as possible to 98% coverage, and (4) where time is not spent learning general words that do not occur in the medical texts.

In the next section, these four criteria will be applied to the med2 corpus using the same existing general and academic word lists (i.e., the GSL, AWL, Pilot Science List, and BNC/COCA lists) and the medical word lists created in previous chapters.

6.5 The behaviour of the existing medical word lists on the med2 corpus

The behaviour of the existing general word lists (i.e., the GSL, and BNC/COCA lists), the academic word lists (i.e., the AWL, and Pilot Science List), and the medical word lists used in Chapters 4 and 5 is now assessed on the independent 5.8 million token medical corpus.

First of all, the lexical profile of the word lists run on the med1 corpus in Chapter 4 is compared with the lexical profile results obtained after running the same lists on the independent medical corpus (i.e., the med2 corpus). Next, the lexical profile of the twenty-five 1,000 BNC/COCA lists and the medical word lists used in Chapter 5 is compared with the lexical profile the same lists have on the med2 corpus. Then, the adequacy of the medical lists on a different medical corpus (the med2 corpus) to the one used for their creation is tested and discussed.

6.5.1 Behaviour of existing medical word lists on the med2 corpus (including the GSL, AWL, and Pilot Science List)

As we can see in Table 6.6, the text coverage results of the GSL1, GSL2, AWL, and Pilot Science List over the med2 corpus confirm that the vocabulary of these four lists includes

words worth knowing by readers of medical texts written in English. (See the detailed coverage results of each word list in Appendix 4.4).

Table 6.6 Occurrence and cumulative coverage of the GSL1, GSL2, AWL, Pilot Science List, the three MGEN lists, and the twenty-three MED lists in the med2 corpus and the med1 corpus

Cumulative results of existing lists		Number of tokens	% of tokens	Number of types	% of types	Number of families
GSL1, GSL2, AWL, Pilot Science List	Med1	4137958	76.18	10841	19.59	2944
	Med2	4491316	76.25	11790	18.15	3007
MGEN (three 1,000) lists	Med1	644827	11.87	3000	5.43	3000
	Med2	668344	11.34	2967	4.56	2967
MED (twenty three 1,000) lists	Med1	540974	9.96	23000	41.63	23000
	Med2	500129	8.49	17099	26.34	17099
Cumulative total of existing lists	Med1	5323759	98.01	36841	66.65	28944
	Med2	5659789	96.08	31856	49.05	23073
Words outside the lists	Med1	107981	1.99	18513	33.44	0
	Med2	230688	3.92	33063	50.93	0
Total	Med1	5431740	100	55354	100	28944
	Med2	5890477	100	64919	100	23073

As summarised in Table 6.6, the GSL1, GSL2, AWL, and Pilot Science List coverage of the med2 corpus is 76.25%. This is virtually the same coverage these same four word lists have on the med 1 corpus (76.18%). When we look at the coverage of these four lists on the med2 corpus separately, the GSL1 and GSL2 have 0.36% and 0.45% higher coverage, while the AWL and Pilot Science List have 0.57% and 0.17% lower coverage, respectively. In fact, this is the only set of word lists in Table 6.6 with a slightly higher coverage on the med2 corpus. In relation to the coverage of the three MGEN lists and the twenty-three MED lists also summarised in Table 6.6, these medical word lists cover 11.34% of the med2 corpus (versus 11.87% of the med1 corpus) and 8.49% (versus 9.96% of the med1 corpus).

Table 6.7 summarises the number of tokens and types in the existing general/academic lists (GSL, AWL, Pilot Science List), and the specialised medical word lists (MGEN, and MED) when run over the med1 and med2 corpora.

Table 6.7 Tokens and word types in the GSL1, GSL2, AWL, Pilot Science List, and medical word lists in the med1 and med2 corpora

Word lists		Number of word types	Cumulative number of word types	Total of word types outside this list	Percentage of tokens in this list	Cumulative percentage of tokens	Percentage of word tokens outside this list
GSL1	Med1	4018	4018	32823	55.34	55.34	42.67
	Med2	4384	4384	27472	55.70	55.70	40.38
GSL2	Med1	2957	6975	29866	6.50	61.84	36.17
	Med2	3365	7749	24107	6.95	62.65	33.43
AWL	Med1	2578	9553	27288	8.28	70.12	27.89
	Med2	2709	10458	21398	7.71	70.36	25.72
Pilot Science List	Med1	1288	10841	26000	6.06	6.18	21.83
	Med2	1332	11790	20066	5.89	76.25	19.83
MGEN lists (three 1,000 medical lists)	Med1	3000	13841	23000	11.87	88.05	9.96
	Med2	2967	14757	17099	11.34	87.59	8.49
MED lists (twenty-three 1,000 only medical lists)	Med1	23000	36841	0	9.96	98.01	0
	Med2	17099	31856	0	8.49	96.01	0

Each of the three MGEN lists has an individual coverage of 5.13% (MGEN1), 3.36% (MGEN2), and 2.25% (MGEN3) (See Appendix 4.4). The 11.34% overall coverage of the MGEN lists is only 0.53% lower on the med2 corpus.

After looking at the behaviour of the newly created medical word lists in Tables 6.6 and 6.7, we notice that the total cumulative coverage figures of running all the word lists through the medical corpora (i.e., 98% coverage of med1 vs. 96% coverage of med2) show a 2% difference in the total coverage between the two corpora. In this respect, based on the results of previous research (Hu & Nation, 2000; Laufer, 1989, 1992; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt et al., 2011) on the most appropriate lexical coverage for unassisted reading comprehension of academic texts (which ranges between 95% and 98% for an adequate and optimal lexical threshold, respectively), we consider that the above mentioned overall coverage results of the validation indicate that the medical word lists function well with both medical corpora (med1 and med2). Additionally, these coverage results demonstrate that the medical word lists created as part of this investigation can be confidently used to obtain reliable estimates of the lexical demands of medical textbooks.

The biggest difference in text coverage between the med1 and med2 corpora is observed in the cumulative coverage of the twenty-three 1,000 MED lists. The cumulative coverage results of these lists show a 1.47% lower coverage on the med2 corpus. Just under half (6.19%) of the total coverage of the MED lists is in the first three MED lists: 4.14% coverage of MED 1st 1,000 list, 1.25% coverage of MED 2nd 1,000 list, and 0.72% coverage of MED 3rd 1,000 list (See Appendix 4.4).

Comparison of the lexical profile results of the med1 and med2 corpora confirms that once readers of medical textbooks are familiar with the words in the GSL, AWL, and Pilot Science List, the most efficient way to sequence the learning of medical words is to focus first on the learning of medical words included in the medical word lists with the highest coverage on medical texts. In the particular case of the medical lists that include words outside the GSL, AWL, and Pilot Science List, it would be ideal for learners of English for Medical Purposes to learn the medical words in the following order: MGEN 1st 1,000 (5.73%), MED 1st 1,000 (4.14%), MGEN 2nd 1,000 (3.36%), MGEN 3rd 1,000 (2.25%), MED 2nd 1,000 (1.24%), and so forth (see Appendix 4.4). That is, it is more efficient to learn the words in the word lists with highest coverage first.

As Tables 6.6 and 6.7 show, the occurrence of word types in the GSL, AWL, Pilot Science List, MGEN lists, and MED lists is in most cases slightly higher in the med2 corpus. This may be attributed at least partly to the additional 9,565 word types in the med2 corpus.

However, when we look at the total number of word types in the twenty-three MED lists, there are 5,901 fewer word types from these MED lists appearing in the med2 corpus. This result is consistent with the lower coverage (8.49%) of the twenty-three MED lists when compared to the coverage results of the same word lists over the med1 corpus (9.96%). This is not too surprising as word lists created from a particular corpus are likely to provide better coverage and have more occurrences in that corpus than in an independent corpus.

Table 6.7 also shows that the GLS1, GSL2, AWL, and Pilot Science List cover a cumulative total of 11,790 word types (949 more word types than in the med1 corpus) and of 20,066 medical word types (5,934 less word types than in the med1 corpus) spread through all the medical word lists used for this study. That is, a total of 31,856 word types in the med2 corpus are needed to get as close as possible to 98%, which was in fact 96.08% for the med2 corpus.

In sum, the results of running all the medical word lists required in Chapter 4 when trying to reach the 98% lexical threshold on a different medical corpus (i.e., the med2 corpus) reveal that 96.08% is the highest lexical text coverage that can be achieved with the existing medical word lists over the new independent (i.e., med2) corpus. (Tables 6.6 and 6.7 are based on data from Appendix 4.4).

6.5.2 Behaviour of existing medical word lists on the med2 corpus (including BNC/COCA lists)

The comparison of the lexical text coverage results of the med1 and med2 corpora using the same BNC/COCA word family lists and medical word lists used in Chapter 5 is now summarised in Table 6.8.

Table 6.8 Number of word types needed to get close to 98% coverage in the med1 and med2 corpora

Assumed vocabulary size (BNC/COCA word family lists used)		Percentage of tokens of assumed vocabulary size	Number of 1,000 medical word type lists outside BNC/COCA lists	Number of word types in the assumed size	Number of word types outside the assumed size	Total number of word types	Total percentage of tokens
3,000	Med1	74.10	27	10345	27000	37345	97.63
	Med2	74.07	27	11386	21112	32498	95.71
4,000	Med1	78.26	16	12440	16000	28440	97.91
	Med2	78.26	16	13764	14709	28473	96.09
5,000	Med1	80.83	12	14110	12000	26110	97.94
	Med2	80.72	12	15665	11310	26975	96.22
6,000	Med1	82.54	11	15503	11000	26503	98.06
	Med2	82.41	11	17236	10561	27797	96.51
9,000	Med1	85.84	9	18242	9000	27228	98.04
	Med2	85.76	9	20469	8689	29158	96.50
10,000	Med1	86.57	9	18841	9000	27841	98.09
	Med2	86.53	9	21226	8635	29861	96.60
25,000	Med1	91.92	6	23867	6000	29853	98.18
	Med2	91.90	6	26998	5723	32721	97.04

Table 6.8 shows the point where, after using the BNC/COCA lists plus the medical lists, the lexical text coverage starts to get close to 98% in the med1 and med2 corpora. In the particular case of the coverage results of the med2 corpus, 97% (96.51%) is the closest to 98% (the optimal lexical text coverage) that the available lists can get on the med2 corpus. The coverage of the med2 corpus first starts getting close to 97% (its highest lexical text coverage figure) when using the first 6,000 BNC/COCA lists plus the twelve medical word lists (see Appendix 4.12 with detailed lexical profile results).

In sum, the point where the smallest number of word types needed to get close to an optimal text coverage (97%) in the med2 corpus requires knowledge of 17,000 words represented by six 1,000 BNC/COCA word family lists and eleven 1,000 medical word type lists. As shown in Table 6.9, the lexical text coverage provided by each subsequent medical word list is consistently higher than the coverage by the following BNC/COCA list. These lexical text coverage results indicate that the earlier the learning of medical words starts, the more efficient the learning of medical words is for the learner of medical English.

We can also see in Table 6.8 that learners of medical English with an average vocabulary size of 3,000 (represented by the first 3,000 BNC/COCA lists) know around 74% of the tokens (i.e., exactly 74.10% of med1 and 74.07% of med2), and if these learners move to learning medical words instead of the 4th 1,000 BNC/COCA list, the next medical word list (i.e., the MGEN 1st 1,000) gives an additional coverage of 8.49% rather than the 4.19% coverage provided by the 4th 1,000 BNC/COCA list over the med2 corpus (see Table 6.9).

Table 6.9 Comparison of coverage provided by the next list of words in the med1 and med2 corpora

Assumed vocabulary size (BNC/COCA word family lists used)		Percentage coverage by the next 1000 BNC/COCA words	Coverage by the next 1000 medical words
3,000	Med1	4.16 (4th 1000)	9.07**
	Med2	4.19	8.49
4,000	Med1	2.57 (5th 1000)	7.17
	Med2	2.46	6.67
5,000	Med1	1.71 (6th 1000)	5.82
	Med2	1.69	5.44
6,000	Med1	1.57 (7th 1000)	5.15
	Med2	1.60	4.48
9,000	Med1	0.73 (10th 1000)	4.86
	Med2	0.77	3.88
10,000	Med1	0.67 (11th 1000)	4.73
	Med2	0.68	3.78
25,000	Med1	Less than 0.6	2.59
	Med2	Less than 0.6	1.85

The lexical text coverage of each subsequent BNC/COCA list is consistently similar in the med1 and med2 corpora. With regard to the coverage of the next lists, these medical lists constantly have a slightly lower coverage of the med2 corpus. This is an expected result considering that the medical lists were created using the med1 corpus.

6.6 Frequency comparisons between the med1 and med2 corpora

The comparison of the lexical profiles of the med1 and med2 corpora has been the main focus of this chapter. Previous analyses in this chapter have shown that there are 55,354 word types in med1 and 64,354 word types in med2. The comparison of the lexical profiles of these two medical corpora has also indicated that 12,191 of the 55,354 word types in the med1 corpus only occur in med1, and 18,447 of the 64,354 word types in the med2 corpus only occur in med2. The comparison in the number of word types in the two medical corpora leaves 40,216 word types occurring both in the med1 and med2 corpora.

Among the most frequent medical words in the med2 corpus, there are words such as *therapy* (fx 7432, rank 76 in the med2 corpus) and *cells* (fx 8301, rank 67 in the med2 corpus). These items do occur in the top ten most frequent medical words in the med2 corpus, but they still are high-frequency words in medical texts. The same happens with medical words in the med1 corpus such as *syndrome* and *symptoms*. In the case of *syndrome* (fx 7981, rank 67 in the med1 corpus) and *symptoms* (fx 7554, rank 62), they are also frequent words with a high-frequency ranking in the med1 corpus.

Table 6.10 compares the frequency of occurrence of the ten most frequent medical words in the two medical corpora.

Table 6.10 The ten most frequent medical words in the med1 and med2 corpora

Rank	Med1 word	Med1 fx	Med1 rank	Med2 word	Med2 fx	Med2 rank
1	patients	37243	14	patients	31947	16
2	disease	22662	19	disease	23735	18
3	treatment	12649	35	treatment	16814	27
4	therapy	11037	43	diagnosis	11338	42
5	infection	10914	45	symptoms	10776	43
6	cells	10870	46	clinical	10731	44
7	clinical	9671	48	infection	10537	47
8	blood	8919	53	syndrome	9826	51
9	diagnosis	8405	58	blood	9251	59
10	acute	8008	61	acute	8523	64

Note in Table 6.10 that eight of the most frequent medical word types in the med1 and med2 corpora are the same. This similarity in the ranking order can be attributed to the fact that the med1 and med2 corpora cover similar topics in the same field (medicine).

Table 6.11 lists the ten most frequent medical word types in the med2 corpus that occur only once in med1. These medical word types are examples of British spelling (i.e., *haemoglobin, ischaemia, coeliac, sulphate, anaesthesia, titres*), names of medical substances in the UK (e.g., *ciclosporin* and *trimozazole*), hyphenated word parts in British English (e.g., *angio-* in words like *angio-oedema*) that form non-hyphenated compounds in American English (e.g., *angioedema*) and derivatives of already existing medical word types and word families (e.g., *histological, pharmacological*).

Table 6.11 Frequency of occurrence of the ten most frequent medical word types in the med2 corpus occurring only once in the med1 corpus

Rank	Word type	Med1 fx	Med2 fx
1	haemoglobin	1	855
2	ischaemia	1	623
3	histological	1	544
4	coeliac	1	329
5	pharmacological	1	285
6	ciclosporin	1	240
7	sulphate	1	199
8	anaesthesia	1	161
9	titres	1	145
10	angio	1	144

In relation to the ten most frequent word types in the med2 corpus that do not occur in med1, most involve British spelling of existing word types in the medical lists (See Table 6.12). The reason why these high-frequency medical word types with British spelling did not make it into the medical word lists used in previous chapters is related to the nature of the texts included in the med1 corpus. That is, the textbooks used in the med1 corpus have been written and published predominantly in the United Stated of America using American English spelling.

Table 6.12 Frequency of occurrence of the ten most frequent medical word types occurring only in the med2 corpus

Rank	Word type	Med1 fx	Med2 fx
1	aetiology	0	1100
2	ischaemic	0	592
3	hypoglycaemia	0	568
4	oesophageal	0	559
5	haemolytic	0	543
6	haemolysis	0	438
7	haematological	0	425
8	haemorrhagic	0	421
9	hypercalcaemia	0	350
10	bacteraemia	0	334

The number of word types across the different frequency bands presented in Table 6.13 shows, in most instances, a higher number of word types in the med2 corpus. This is to be expected, since the med2 corpus is bigger than the med1 corpus and contains 458,733 more tokens and 9,565 more word types. However, the percentage of word types across the different frequency bands in Table 6.13 is in general very similar in both medical corpora, being in most cases no more than 1% higher or lower in the med2 corpus (e.g., 29.66% of word types occurring once in med1 vs. 30.93% occurring once in med2).

Table 6.13 Number of word types occurring at different frequency levels

Occurrences	Med1 corpus	Med1%	Med2 corpus	Med2%
Once	16417	29.66	20080	30.93
Twice	6634	11.98	8214	12.65
3 to 9 times	13586	24.54	15982	24.62
10 to 19 times	5520	9.97	6207	9.56
20 to 49 times	5240	9.47	5814	8.96
50 to 99 times	2822	5.10	3123	4.81
100 to 299 times	2831	5.11	3079	4.74
300 to 499 times	852	1.54	852	1.31
500 to 999 times	725	1.31	778	1.20
1,000 to 9,999 times	682	1.23	742	1.14
10,000 to 99,999 times	40	0.07	42	0.06
100,000 or more	5	0.01	6	0.01
Total	55354	100	64919	100

Table 6.13 shows the distribution of frequencies across the med1 and med2 corpora. When we look at the number of word types within each frequency band, the higher the frequency of occurrence of a word type, the smaller the gap between the number of word types occurring in the med1 and med2 corpora. On the other hand, we can also observe in Table 6.13 that while the percentage (0.01%) of word types occurring 100,000 times or more is almost the same in both medical corpora, the percentage of word types occurring only once is no more than 1.22% higher in med2. In short, the percentage of medical word types in the two medical corpora gets closer when we get to the word types with the highest frequencies. A similar general trend (but with a bigger frequency gap) was noticed when comparing the number of occurrences in the med1 and general corpora of the same size in Chapter 4.

6.7 Conclusions

The lexical profile results of Chapters 4, 5 and 6, and the coverage results of running the medical word lists on the med2 corpus suggest the following:

1. The lexical profile results obtained after running the med1 corpus and the med2 corpus separately through various existing general and academic word lists, and the medical word lists show very similar text coverage of both corpora (i.e, 98% coverage of med1 vs. 96% coverage of med2). The biggest difference in coverage is observed in the 2% overall lower coverage of the med2 corpus over the word lists. This lower 2% coverage can be attributed mainly to the use of the medical lists created using the med1 corpus on a different corpus (i.e., the med2 corpus). Moreover, if alternative British spelling would have been considered when creating the new sets of medical word lists, since only medical word types with American spelling were included, the overall 2% higher coverage of the medical lists over the med1 corpus may have been smaller.
2. The sooner medical students start learning the medical words related to the medical field, the more efficient the acquisition of the specialised (medical) vocabulary they need to learn.

Chapter 7: Discussion of the findings

The present thesis has explored two aspects of needs analysis (Hutchinson & Waters, 1987) in ESP – lacks (i.e., vocabulary size) and necessities (i.e., lexical demands of medical texts). As the starting point of the analysis, the first study in this thesis (see Chapter 3) estimated the vocabulary size of a group of 408 ESP learners in two Spanish medium higher education institutions in Venezuela. The VST – the instrument selected for this purpose – was used as a diagnostic measure of the ESP learners' lexical knowledge of general English at the beginning of their first ESP course at university. Then, in Chapters 4, 5 and 6, this research proposed a methodology to identify and validate the salient lexicon of written medical texts. In the present chapter, some useful connections between the lacks (vocabulary size) and necessities (medical vocabulary) strands of this investigation and its findings are established. These connections are discussed in relation to three aspects of the study: the approach followed for the identification of medical words, the ESP background of the study, and the implications of the findings.

7.1 The approach to the identification of medical vocabulary: challenges and value

The medical vocabulary strand of this investigation replicated Chung and Nation's (2003, 2004) methodology for identifying content area (technical) words. As previously discussed in Chapters 4 and 5 their methodology is twofold, involving the use of a semantic rating scale, and a corpus comparison approach. Chung and Nation's methodology was also the approach used in this thesis to quantify the lexical demands of reading medical textbooks. The overall lexical coverage results of the present investigation showed that 36.47% of the running words (tokens) in the medical textbooks were medical words. A coverage result that is 5% higher than Chung and Nation's (2003) coverage figure of 31.2% of running words for anatomy texts. The 36.47% coverage of running words for texts of general medicine indicates that the use of Chung and Nation's approach for the present investigation is an effective way for identifying discipline-specific vocabulary.

7.1.1 The challenges and value of the approach to the identification of medical vocabulary

The challenges and value of the twofold methodology here adopted for identifying medical words are discussed below. This discussion revolves around the following aspects of the present investigation: the semantic rating scale, the corpus comparison approach, the size of the corpus, the size of the medical word lists, and the validation procedure for the newly created medical word lists.

7.1.1.1 The semantic rating scale

The replication of Chung and Nation's (2003) semantic rating scale involved the identification of thousands (over 30,000) of medical words occurring in the medical corpus used for creating the lists, i.e., the med1 corpus. In spite of the usefulness of the semantic rating scale for making decisions on the number of content area vocabulary found in medical texts, the need to classify such a large number of words made the use of the rating scale a very time-consuming process (as also reported by Chung & Nation, 2004; Fraser, 2005, 2006). Thus, the three main reasons that delayed the identification of medical words while using the rating scale were as follows: (1) the large number of medical words (32,194 word types) classified as having a medical meaning, (2) the use of a specialised medical dictionary to look up words with unknown or unclear meaning, and (3) the use of concordance lines to look at the words in their context of occurrence. These lines were, however, useful when in doubt of where to place the words on the rating scale based on how closely related to the medical field a particular words was. In general, an average of 2,500 medical types were classified per week using the semantic rating scale, and the overall identification of the 32,194 medical word types took over three months. Likewise, there were still over 8,000 word types, most of them words occurring only once, that remained unclassified mainly because these words were not needed to estimate the 98% lexical threshold.

The adaptation of Chung and Nation's (2003) rating scale for the present investigation, as discussed in Chapter 4, has enabled us to provide a comprehensive account of the lexical demands of medical textbooks from a semantic perspective. 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 med1 corpus – occurring in existing word lists such as West's

(1953) GSL, Coxhead's (2000) AWL, Coxhead and Hirsh's (2007) Pilot Science List, and Nation's (2012) BNC/COCA lists, and beyond those word lists. Overall, the use of Chung and Nation's semantic rating scale enabled the identification of an enormous number of content area (technical) words (36.47%) found in medical texts written in English.

7.1.1.2 The corpus comparison approach

Initial piloting of the corpus comparison approach on a medical corpus of around 500,000 tokens showed that a larger corpus was required to enable the quantification of the 98% lexical threshold of medical textbooks. Additionally, the piloting revealed that simply ranking the word types occurring in the medical corpus by frequency was not enough to get a clear picture of the lexical profile of medical textbooks. Thus, the results of the piloting indicated that a larger corpus and a different procedure for sorting the medical words were required to create the medical word lists needed to estimate the vocabulary load of medical textbooks.

As discussed in Chapter 4, two larger corpora (i.e., a medical corpus and a general corpus) than the 500,000 token medical corpus created for the piloting were compiled to enable the implementation of the corpus comparison approach. These two larger corpora were characterised by having the same size, 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 (the med1 corpus) was created for identifying medical vocabulary, using Chung and Nation's semantic rating scale, in well-known existing word lists – such as West's (1953) GSL, Coxhead's (2000) AWL, Coxhead and Hirsh's (2007) Pilot Science List, and Nation's (2012) BNC/COCA lists – 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 texts written in English.

7.1.1.3 The size of the corpus

In relation to the size of the medical corpus, the results of piloting a medical corpus of 500,000 tokens, created prior to compiling the med1 corpus, justified our decision to increase the size of the medical corpus from 500,000 to 5 million tokens. As mentioned in

the previous section, the piloting revealed that a specialised corpus of 500,000 tokens did not include enough instances of words to enable a reliable estimation of the vocabulary load of medical texts. Consequently, the size of the med1 corpus was determined ultimately by the amount of specialised texts from a wide range of medical topics available in digital format. This also led in the end to the compilation of a medical corpus ten times larger than the medical corpus originally compiled for the piloting. A ten times larger medical corpus meant that a wider range of medical topics and words were likely to occur in the med1 corpus, thus facilitating the estimation of the lexical demands of medical textbooks. In fact, the size of the med1 corpus was large enough in number of running words (5,431,740 tokens) and coverage of medical topics to provide a representative sample of the lexis found in medical textbooks.

The use of two corpora (i.e., the med1 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 thesis. As shown by the lexical coverage results of these two corpora, the total number of word types in the medical corpus is nearly half the size (55,354 word types in the medical corpus vs. 97,648 word types in the general corpus). Also, 32,194 of those word types occurring in the medical corpus have been identified as having a medical meaning. Of those 32,194 medical words, over 23,000 are unique to the medical corpus. This is a clear indication that not all the vocabulary occurring in the medical corpus can be found in the general corpus, confirming the specialist nature of much of medical vocabulary. These figures emphasise the restricted nature of the vocabulary occurring in the medical texts examined in the present study. (See Chapters 4 and 5 for further discussion of the findings of the corpus comparison approach).

7.1.1.4 The size of the medical word lists

The creation of a series of medical word lists, using the above mentioned twofold methodology, has made possible the quantification 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 content area

(medical) vocabulary as early as possible, probably not before studying and becoming familiar with specialised subject content, but if possible at the same time.

First of all, over 30,000 of the most frequent word types in the med1 corpus were classified as medical words using an adaptation of Chung and Nation's (2003) semantic rating scale, as explained in Chapter 4. Then, the words identified as having a medical meaning were grouped into medical word lists of 1,000 word types. Two different frequency-based procedures were used to rank and group the medical words previously classified.

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 Pilot Science list (see Chapter 4 and Appendix 2.1), and four 1,000 MGEN word lists beyond the first three 1,000 BNC/COCA word lists (see Chapter 5 and Appendix 3.8) 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 (see Chapter 4), 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 frequency-based procedure, it included medical words that only occurred in the medical corpus, that is, words unique to the medical corpus. Following the second procedure, the medical word types were ranked by their highest absolute frequency of occurrence in the med1 corpus. A total of twenty-six 1,000 medical word lists were created, including words beyond the GLS, AWL, and Pilot Science list in Chapter 4, and beyond the BNC/COCA lists in Chapter 5.

The use of Chung and Nation's (2003) 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 a number of purposes. For example, 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 English for medical purposes lesson, 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.

7.1.1.5 The validation with another corpus

Apart from compiling the two corpora (i.e., the med1 and general corpora) above mentioned, a third corpus (the med2 corpus) was created for validating the series of newly created medical word lists already discussed in Chapters 4 and 5. That is, a second medical corpus (the med2 corpus) of similar size (5,890,477 tokens in med2 vs. 5,431,740 tokens in med1) to the med1 corpus was compiled. (See Chapter 6 for further details). The validation of newly created word lists using another corpus has been a standard methodological practice in several studies (e.g., Coxhead, 2000; Coxhead & Hirsh, 2007; Fraser, 2007; Konstantakis, 2010) that have reported on the creation of academic and specialised word lists.

The med2 corpus was created taking into account important features of the med1 corpus, such as the size of the corpus, range of health and medical topics included, and written nature of the texts selected. In general, the compilation process of the med2 corpus involved careful comparison with the features of the med1 corpus. The main purpose of the med2 corpus was to investigate the behaviour of the newly created medical word lists on a different medical corpus (the med2 corpus), and check their reliability on an independent medical corpus, namely the med2 corpus.

The validation of the series of medical word lists was conducted by estimating the coverage of the newly created medical word lists using the independent medical corpus (the med2 corpus) above mentioned which consists of different complete chapters from medical textbooks (see Chapter 5) and is of a similar size (5,890,477 tokens) to the med1 corpus – the medical corpus originally compiled to create the medical word lists. Lexical profile results showed very similar text coverage for both medical corpora (i.e., 98% coverage of med1 vs. 96% coverage of med2). These coverage results also indicated the medical word lists created as part of this investigation can be confidently used to obtain reliable estimates of the lexical demands of medical texts.

These three corpora (i.e., the med1, med2 and general corpora) are of great value to apply the methodology used in this thesis to estimate the lexical demands of medical textbooks and validate the series of newly created medical word lists. While the med1 and general corpora have been useful for identifying medical lexis, the med2 corpus has made possible the validation of the series of medical word lists created for the present investigation.

In short, when referring to both the challenges and value of using Chung and Nation's (2003) methodology to create a series of medical word lists and to investigate the lexical profile of medical texts, the comprehensive methodology of the present research represents an improvement over recent studies (Fraser, 2007, 2009, 2013; Fraser et al., 2015; Hsu, 2013; J. Wang et al., 2008) specially in terms of the large number of content specific (medical) words classified, the size of newly created medical word lists, and the validation of the medical lists created as part of this investigation on the vocabulary load of medical textbooks.

7.2 An ESP course

As previously mentioned in Chapters 1 and 2 of this thesis, the ESP teacher needs to be familiar with the various levels of linguistic description and analysis (i.e., phonetic and phonological, morphological, lexical, syntactic, semantic, and pragmatic) and language variation (i.e., lexicon, discourse, and dialect) that characterise a given discipline or content area. In the case of the ESP teacher for medical students, this particular group of language educators needs to be familiar with the linguistic features of the language variety or occupational dialect (as referred to by Millward & Hayes, 2012) of the medical discourse.

With respect to knowledge of the disciplinary content, the ESP teacher is first in a much stronger position, by being the main source of information. The ESP teacher in the Venezuelan context deals with the content of the readings by planning and promoting pre-reading, reading and post-reading activities that encourage ESP learners to (1) share, in their L1, their background knowledge of a given specialised topic or theme (e.g., as part of the pre-reading activities and group discussions), (2) extract key information contained in the readings, (3) make meaningful connections between the previous information and new information (e.g., as part of the reading and post-reading activities) related to the specialised topics discussed in the ESP reading classes. Particularly in the first half of the ESP course,

the ESP teacher provides L2 strategy training, highlights the salient linguistic features of the specialised texts discussed, and plays a leading role in the selection of the content area readings to be used in the ESP classes.

As the ESP course progresses, the students are in a much stronger position and have gained in confidence by: (1) participating more actively in the selection of the relevant specialised readings for class discussion in the ESP classes, (2) being more motivated to take part in group work, (3) being more willing to give short in-class presentations on discipline-specific topics of their choice, and (4) being more confident users of the L2 learning strategies developed through the ESP course.

Most ESP students in Venezuela begin to learn about their subject areas in their L1 (Spanish) from the start of their undergraduate degrees, while simultaneously recycling and discussing some of the topics of their discipline-specific subjects in the ESP reading classes. That is, ESP learners in Venezuela are likely to be reading, at the same time, about similar disciplinary content in their L1 (i.e., Spanish in their discipline-specific subjects) and their L2 (i.e., English in the ESP reading classes). In the particular case of undergraduate medical students at the start their first ESP reading course, this disciplinary content knowledge comes mainly from the Natural Sciences, and Health Sciences subjects they became familiar with throughout their secondary school education. Since their disciplinary content knowledge is not well established yet in their L1, particularly at the beginning of their first ESP course, then dealing with medical texts in an L2 (i.e., English) is even more challenging from a disciplinary knowledge perspective.

The selected readings for the ESP course for medical undergraduates come mostly from authentic materials (e.g., reference books, textbooks, and specialised periodicals, among others) and refer to topics from their content areas of specialisation. In most cases, the topics of the ESP reading materials are closely related to the specialised topics the students are simultaneously learning about in the discipline-specific subjects. For example, the ESP readings for the medical and health sciences undergraduates in Venezuela come from discipline-specific extracts, chapters, sections, or articles taken mainly from medical and scientific reference books, textbooks, journals and magazines (e.g., Discover, Scientific American, Nature, Science, Science Daily, and The Scientist, among others).

The results of the lacks aspect (i.e., vocabulary size) of needs analysis, already discussed in Chapter 3, revealed that this group of L1 Spanish test takers has an average vocabulary size of 6,300 word families. Moreover, when we apply these results to the particular population of ESP learners who are about to start their first English for Medical Purposes reading course at university, we can conclude that prospective medical and health sciences undergraduates in Venezuela know enough general vocabulary in English to begin their study of reading in English for Medical Purposes. However, as seen in Table 7.1, these 6,300 words are not all from the first six 1,000 word frequency bands. To illustrate this idea further, the percentage of known words per frequency level of the 14,000 version of the VST, as well as the spread of the average number of correct responses across the 14 frequency levels sat by the ESP test takers in Venezuela are summarised in Table 7.1 and Figure 7.1.

Table 7.1 Average of VST tested words known by ESP test takers

Frequency band	Percentage (%) of correct responses	Number of cognates	Number of false cognates	Total number of target words per level
Level 1	59	5	1	10
Level 2	49	4	1	10
Level 3	47	4	0	10
Level 4	49	6	1	10
Level 5	49	7	0	10
Level 6	47	6	0	10
Level 7	42	4	1	10
Level 8	54	7	0	10
Level 9	43	5	0	10
Level 10	35	2	0	10
Level 11	47	6	0	10
Level 12	41	6	0	10
Level 13	37	7	1	10
Level 14	44	7	1	10

As shown in Table 7.1, there is a general tendency across the 14 frequency levels to have, on average, a smaller number of correct answers at the subsequent lower frequency level. That is, the lower the frequency level the smaller the average percentage of target words the VST test takers know. This decreasing tendency in the number of tested words known is clearly observed at levels 2, 3, 4, 7 and 10. However, this downward tendency in the percentage of correct answers is less obvious at other frequency levels (e.g., levels 5, 8, 11 and 14) where over 50% of the tested words are identical or nearly identical English/Spanish cognates or loan words.

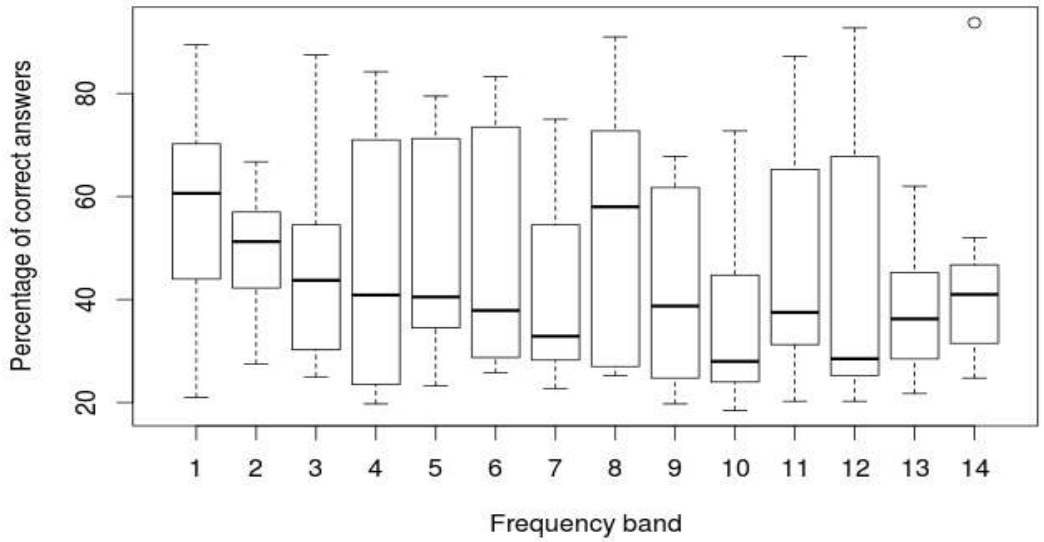

Figure 7.1 Box plot for the mean correct answers by frequency band for the VST taken by Venezuelan undergraduate students. Notice the outlier on band 14. This corresponds to the word '*cordillera*'.

As seen in Table 7.1, the analysis of the number of correct answers by frequency band reveals that the number of correct answers tended to decrease as the word frequency was lower. An exception to this tendency is observed in bands 8 and 11. The higher average of correct answers in these bands is largely due to English/Spanish cognates. In other words, the English/Spanish cognates included in the test had an important effect on the number of correct answers. Not surprisingly, there was a higher number of correct responses for cognates than for non-cognates. Consequently, the average of correct answers for a particular frequency band was higher when the questions for this band included a majority of English/Spanish cognates. This effect of cognates can be observed in Figure 7.1. Band 8 is shown to be particularly affected by cognates such as *'authentic'*, '*cabaret*', '*eclipse*', '*erratic*', '*kindergarten*', '*null*', and '*palette*'. We suggest that the word frequency of the cognate in Spanish also has an effect on Spanish speakers to recognise English words. For example, the words '*kindergarten*', '*authentic*', '*puma*', '*yoga*' and '*reptile*' are, not only transparent cognates, but frequent words in Spanish. On the other hand, '*azalea*', '*regent*', and '*refectory*' are low-frequency words in Spanish.

Both Table 7.1 and Figure 7.1 demonstrate that the words prospective ESP undergraduates in Venezuela know are distributed along the 14 frequency bands. According to Nation and Coxhead (2014), the test takers' ability to recognise words, even at the lowest frequency level (i.e., the fourteen 1,000 frequency band), can be attributed to several factors, such as their prior knowledge, ability to remember words from previous foreign language learning experience in English or other foreign languages, their pastimes, and interests, among other educational or life experiences the test takers may have had. For example, the knowledge by most of test takers of the outliner and identical English/Spanish cognate '*cordillera*' at frequency band 14 (see Figure 7.1) is linked to the life experiences of the test takers whose universities are located in the '*cordillera*' (i.e., the Andes mountain range).

As L1 speakers of Spanish (i.e., a language with the same Latin script used in English and a good proportion of shared English/Spanish cognates), ESP learners in Venezuela are likely to know English words even at the lowest frequency bands of the VST. For this particular group of test takers, it is important to take into account Nguyen and Nation's (2011) suggestion on the importance of sitting the whole VST or, at least, a version of the test that includes a representative sample of target words for all the levels of the test.

Even though, this research was conducted with a specific group of ESP learners in mind, i.e., prospective undergraduate ESP learners (in Chapter 3) and English for Medical Purposes learners (in Chapters 4, 5, and 6) in ESP reading programmes in Venezuela; the methodology followed to create discipline-specific lists in general (i.e. words lists of frequent and salient vocabulary in health and medical sciences textbooks), and the content area (medical) lists in particular developed in Chapters 4 and 5 can be useful for a wider audience of ESP and EMP teachers and learners. For example, ESP teachers of medical and health sciences in various parts of the world – such as Iran (Ghalandari & Talebinejad, 2012), Romania (Popa, 2013), Saudi Arabia (Javid, 2011), and Pakistan (Niazi, 2014), among many others – can benefit from using the medical lists created in the present investigation. These medical word lists can be used to determine the vocabulary worth focusing on when teaching the vocabulary learning strategies English for Medical Purposes students need to independently continue with the acquisition of the discipline-specific lexis.

Next, some vocabulary learning strategies worth focusing on in an ESP reading course for undergraduate medical students are suggested in the implications of the findings section.

7.3 Implications of the findings

As demonstrated by the lexical profile results presented in this thesis, the large amount of content area (medical) vocabulary to learn – i.e., around 12,000 new word types beyond the average 6,000 word families known by ESP learners who sat the VST in Venezuela – stresses the importance of devising a plan for ESP reading courses. In descending order of importance, the findings of the present investigation on the huge number of medical words to learn underline the value of (1) strategy training in the ESP reading courses for medical students, (2) learning medical vocabulary as early as possible, and (3) having a reasonable vocabulary size before starting medical study and testing the vocabulary size of ESP learners.

7.3.1 Strategy training in ESP classes for medical students

When planning the vocabulary component of an ESP reading course, the teacher should prioritise spending class time on strategy training. For Nation (2013, p. 342), strategy training plays an important role in second language vocabulary development. Two main purposes of having a strategy training plan are to provide ESP learners with a repertoire of activities they can draw from, at their convenience, based on their lexical needs and particular learning styles, and to equip L2 learners with the linguistic tools to learn how to manage their own language learning and become autonomous readers. As part of the strategy training plan, the ESP teacher should promote vocabulary learning activities that allow students to embark on independent vocabulary learning activities both inside and outside the ESP classroom. The acquisition of the vocabulary learning strategies suitable to the specific learning needs and styles of the ESP learners is what will eventually enable medical students to cope with the lexical demands of medical texts and achieve the 98% lexical threshold required for appropriate reading comprehension of such specialised texts.

In the ESP reading courses for medical students, strategy training should focus on the development of the following vocabulary learning strategies: the general strategy of word consciousness, the word part strategy, the word card strategy, and the strategy of using in-text definitions and visual representations.

7.3.1.1 The general strategy of word consciousness

The strategy of word consciousness involves promoting language learning tasks that develop, according to Scott and Nagy's (2004) definition of the concept, an awareness of and interest in the form, meaning and use of words. A cognitive and an affective component are two key features normally associated with the construct of word consciousness (Anderson & Nagy, 1992; Graves & Watts-Taffe, 2008, 2002; Scott & Nagy, 2004) in first language acquisition.

Word consciousness can take place at different linguistic levels: phonemic (word decoding), morphological (word-part analysis), semantic (word meaning), syntactic (word order), and pragmatic (word use in context). Consciousness raising tasks that focus the L2 learner's attention on specific forms at various linguistic levels play an important role in the process of second language acquisition. These consciousness raising tasks promote noticing and facilitate language learning (Ellis, 2003; Rutherford & Sharwood Smith, 1985; Schmidt, 1992; Willis & Willis, 1996) in general and L2 vocabulary acquisition (Nation, 2013) in particular.

As part of the ESP teacher's crusade to raise his/her students' consciousness and motivation to learn new words, the ESP teacher needs to make sure L2 learners have both components. In this respect, activities that (1) foster awareness of the importance of vocabulary learning strategies, (2) make use of cognates, (3) connect known uses of technical words to their technical uses, (4) encourage the L2 learner to develop a customised plan for learning vocabulary, and (5) make students aware of the size of the vocabulary learning task should be promoted in the ESP classroom. Let us make some instructional considerations on the use of these five word consciousness-raising activities.

Awareness of importance of vocabulary learning strategies. According to Nation and Gu (2007, p. 82) , the implementation of vocabulary learning strategies should consider the individual learner's characteristics, the demands of the task, and the contextual constraints of the teaching and learning situation. On the one hand, the ESP teacher should be aware of the crucial role s/he can play in the development of the appropriate vocabulary learning strategies by the ESP learners and the design and implementation of a suitable strategy training plan that can be integrated into the curriculum of the ESP course. In the case of a

reading course for ESP medical students, for example, a way to reduce the learning burden of content area (medical) vocabulary can be designing a vocabulary training plan that enables the teaching of the vocabulary learning strategies the medical students need to become fluent readers of medical textbooks, and play an active role in the acquisition of medical vocabulary from their specific areas of specialisation.

On the other hand, the ESP learners in the Venezuelan context, for instance, are not always aware of the cognitive tasks involved in learning to read in a foreign language in general, let alone of the lexical demands posed by discipline-specific texts in particular. The learner's awareness of the importance of vocabulary learning strategies could be raised by: informing the ESP learners in class of the specific linguistic features of the specialised language from their particular discipline, promoting activities that familiarise the learners with the appropriate vocabulary learning strategies, involving the students in their own vocabulary learning, and making students take control of the tasks (e.g., planning, selecting, implementing, monitoring, and evaluating) involved in learning vocabulary in another language. (See Gu, 2005 for further discussion on L2 vocabulary learning strategies). Therefore, it is important for both the ESP teacher and learner to be aware of the key role played by strategy training as part of the process of learning vocabulary in another language.

The use of cognates. Cognate-awareness, as a valuable word-consciousness strategy, should be definitely promoted from the very start of an ESP reading programme. Medical vocabulary in English is well-known for having borrowed many terms from Latin and Greek (Salager-Meyer, 1985, 2014) and sharing an important number of cognates with Romance languages like Spanish makes it useful for the ESP teacher to draw the attention of the medical students to words of similar form and meaning. The value of developing cognate-awareness in ESP learners lies in the fact that recognition of English/Spanish cognates with formal variations in spelling patterns (e.g., '*cell'* in English vs. '*célula'* in Spanish, '*dehydration'* in English vs. '*deshidratación'* in Spanish, and '*chemotherapy'* in English vs. '*quimioterapia'* in Spanish) is not always as quick and straightforward for ESP leaners with L1 Spanish as we may assume.

The connections between known and unknown uses of technical words. Some word consciousness-raising tasks can raise ESP learners' awareness of the core meaning of

known words with a general and a medical usage (e.g., 'the state of an object' as a core meaning for the word *condition*; 'a point at which two or more things are joined' as a core meaning for the word '*joint'*; and 'arduous work, effort or task' as a core meaning for the word '*labour'*). Additionally, the core meaning of the words *condition, joint,* and *labour* can be expanded to more subject-specific uses of the same word (i.e., 'a state of health' as a medical meaning for the word '*condition'*, 'union of two or more bones' as a medical meaning for '*joint'*, and 'the process of expulsion of a fetus' as a medical meaning for '*labour'*). In this respect, the ESP teacher can also promote the development of word consciousness through activities that allow extending known meanings and uses of words to unfamiliar (e.g., undiscovered, hidden or more specialised) uses of the same word forms (Nation, 2013, p. 106).

A plan for learning vocabulary. There should be a vocabulary learning plan tailored to the specific learning needs of the ESP students. For example, a customised vocabulary learning plan for undergraduate medical students in Venezuela should be suitable to their learning styles and needs. This plan should enable them to continue to independently acquire the specialised lexis they will need, even after completing their ESP reading courses, for gaining adequate reading comprehension of medical texts written in English. It is important that the ESP teacher guides the learners in the design of a plan that sets both short-term and long-term vocabulary learning goals. Hence, as part of an ongoing process of learning medical vocabulary in a foreign language, it is crucial to equip ESP medical students with the strategy training needed for them to be able and motivated to adapt, when necessary, their own plan for learning medical vocabulary.

Awareness of the size of the vocabulary learning task. Ideally, the language teacher should be informed and willing to raise his/her students' awareness of the lexical demands posed by texts from the specific content area s/he is teaching. In the case of the teacher of an ESP reading course, which is the case of the researcher of the present investigation, lexical profile information for specific subject areas could be obtained from various sources, such as (1) vocabulary research – published in academic journals, websites, blogs, wikis, etc. – reporting on the size of the vocabulary learning task for specific disciplines, (2) ESP course books developed using a frequency-based approach to present, use and recycle content area vocabulary, and (3) subject-specific word lists that can provide a lexical shortcut for guiding vocabulary strategy training as well as for accelerating the learning and growth of

ESP vocabulary. Nonetheless, to the best of the researcher's knowledge, there is still a great need for more specialised vocabulary research that uses a comprehensive approach to investigate the vocabulary load of texts from specific content areas.

Overall, the general strategy of word consciousness highlights the value of promoting word awareness and building lasting motivation in the ESP learners so that they can continue to learn independently the subject-specific vocabulary required for appropriate reading comprehension of specialised texts, and for keeping up to date with the latest scientific discoveries published in English.

7.3.1.2 The word part strategy

The ESP teacher can help the medical students focus, as part of his/her strategy training plan, their attention on word part analysis and frequent affixes and stems worth learning in medical English. The development of word-part awareness of medical vocabulary will increase the learner's ability to relate the meaning of known affixes and stems, to new words (Nation 2013). For example, if the learners are familiar with word parts like the affixes *ab-* (absent, away), *pre-* (before), *-ous* (possessing), and *-ity* (denoting state or condition), and the stems *cancer* (tumour), and *normal* (usual, typical), they are likely to understand the meaning of previously unmet words like *abnormality* and *precancerous.*

7.3.1.3 The word card learning strategy

Noting unknown words from the medical readings onto bilingual word cards or flashcards helps focus the learners' attention on new vocabulary and promotes deliberate learning (Nation, 2013, Chapter 11). Word cards, for instance, can be promoted among ESP learners for self-study of technical terms, and self-evaluation of the learning of subject-specific words. They can also be used for the purpose of memorising factual information about those terms. The ongoing use of word cards including subject-specific vocabulary from the medical reading encourages noticing and learning of words from a particular content area.

7.3.1.4 The strategy of using in-text definitions and visual representations

Another useful vocabulary learning strategy includes vocabulary learning activities that provide opportunities to meet and practice relevant words embedded in definitions and a variety of visual representations such as pictures, tables, diagrams, figures, graphs and so on. In medical texts many of technical terms are defined in the text. Learners should be skilful in dealing with these pieces of information. In addition, quite a few words are in labelled diagrams which provide in effect the meaning of the words. Learners should give attention to these clues (Bramki & Williams, 1984).

Undoubtedly, deliberate word learning through the use of adequate vocabulary strategies (e.g., the general strategy of word consciousness, the word part strategy, the word card strategy, and the strategy of using in-text definitions and visual representations above mentioned) should be taught and promoted among ESP learners.

7.3.2 The importance of learning subject-specific vocabulary

Another implication of the findings of the present research is related to the value of learning medical vocabulary as early as possible. The results of this study have shown how a large proportion of the running words in a medical text are technical vocabulary.

Regardless of their L1, readers of medical textbooks need to know between 26,000 and 27,000 medical word types beyond (the top 3,000 most frequent words) existing word lists – as represented by the GSL, AWL, and Pilot Science list or the first three 1,000 BNC/COCA lists, respectively – to be able to meet the lexical demands of medical textbooks. (See Chapters 4 and 5 for further discussion). However, the learning of, at least, 26,000 medical word types should not be the aim of the vocabulary instruction component of an English for Medical Purposes reading course at university level. The medical students are expected to learn these 26,000 subject-specific words over time. The results of the identification of medical lexis demonstrate the importance of starting learning content specific vocabulary as early as possible.

7.3.3 The ideal vocabulary size for ESP learners before starting medical study

A third implication of the findings of the present investigation refers to the value of having an adequate vocabulary size, testing the vocabulary size of medical students in an ESP teaching and learning context, and using subject-specific tests. In relation to the most suitable point to start the strategy training of medical students, fortunately the results of the administration of the vocabulary size discussed in Chapter 3 indicate that most ESP learners who sat the VST in Venezuela have a reasonable vocabulary size at the start of their first ESP reading course to begin the strategy training required to be familiar with the vocabulary learning strategies and develop language learning autonomy. These results suggest a good start to the vocabulary learning of content area vocabulary for medical students in particular is when they have an average vocabulary size of around 5,000 word families in general English – see Chapter 5. Moreover, the VST results justify the move towards content specific vocabulary at earlier stages of the vocabulary learning process for medical students.

In order to have a clearer picture of the lacks (i.e., vocabulary size) aspect of needs analysis in a particular group of L2 learners, like the Venezuelan ESP learners in general and the medical students in particular who served as the initial motivation for this investigation, it is undoubtedly useful for both ESP teachers and learners to have an estimate of the vocabulary size of a specific group of L2 learners at different stages of the L2 vocabulary acquisition process. The estimation of the vocabulary size of ESP medical students in Venezuela, for instance, can guide the ESP teacher in the planning of the most effective vocabulary strategy training for undergraduate students enrolled in an English for Medical Purposes reading programme. Additionally, if a group of ESP medical students is aware of their vocabulary size at the beginning of their first English for Medical Purposes reading course, this awareness is also a good starting point to develop their consciousness of where they stand in terms of their current general vocabulary size in English, and where to start learning content area (medical) vocabulary.

7.4 Conclusions

The results of the two aspects of needs analysis discussed in this chapter, namely lacks (the average vocabulary size of around 6,000 word families of 408 ESP learners), and

necessities (the need of knowing at least 26,000 medical word types for reaching the lexical threshold of medical textbooks) indicate that an ESP reading course for medical undergraduates in Venezuela could not possibly teach more than a small proportion of the specialised vocabulary required for reaching the 98% lexical threshold. The large number of medical words (i.e., between 26,000 and 27,000 medical word types) that need to be learnt makes it unrealistic to expect that an ESP reading course for undergraduate medical students in this context could focus on teaching all this vocabulary, but instead the above mentioned vocabulary load results of medical texts make it imperative to focus on the development of the vocabulary learning strategies needed by this group of medical students to be able to comprehend medical textbooks written in English.

It would nevertheless appear worthwhile to reiterate here that an ESP reading course for medical students should focus on vocabulary strategy training. An English for Medical Purposes reading course should, therefore, support the learners' medical studies by providing them with the L2 foundations to be able to: (1) comprehend medical texts written in English, and (2) consult a variety of discipline-specific written materials (e.g., periodicals, books, dictionaries, encyclopaedias, scientific records or databases, and medical diagnostic websites, among others) to keep up to date with the discoveries, new information published in the field for their discipline-specific course assignments, and also with their future professional development and research in their field of specialisation.

Chapter 8: Conclusions, pedagogical discussion, limitations and further research

This final chapter begins by briefly recapitulating the major outcomes of previous chapters (i.e., Chapters 3, 4, 5 and 6). This is followed by a pedagogical discussion of the key aspects that English for Medical Purposes teachers need to consider when planning the vocabulary component of a reading course for L2 English for Medical Purposes students. This discussion revolves around the four main roles of the L2 teacher proposed by Nation (2008), namely, planning, strategy training, testing and teaching vocabulary. Once the instructional implications of the present study have been addressed, some comments on the perceived limitations of key topics related to this investigation (e.g. vocabulary size testing, tagging of medical corpora, etcetera) are mentioned. Finally, some ideas for future research are presented before making some closing remarks.

8.1 Conclusions

Here we reflect on the most prominent findings of this thesis in relation to the vocabulary size of the group of ESP learners tested in Chapter 3, the lexical coverage results obtained in Chapters 4 and 5, and the validation of the medical word lists discussed in Chapter 6 of the present study:

1. *Vocabulary size.* The results of measuring the vocabulary size of 408 EFL learners who were prospective students of ESP at two Spanish medium higher education institutions in Venezuela indicated that the average vocabulary size of these L1 Spanish speakers ranges between 6,000-7,000 word families. Thus, we may confidently conclude that this group of ESP learners knows an average of 6,000 word families at the start of their first ESP course at university. Nevertheless, we do not know exactly which words they know or whether all these 6,000 words belong to the first 6,000 BNC/COCA frequency levels. In relation to the percentage of medical words in these first BNC/COCA levels, medical words make up 14.63% of the first 6,000 word families of the BNC/COCA. These families include 4,712 medical word types. (For further details on the spread of medical words across the

BNC/COCA lists, see Appendix 3.5). A vocabulary size of 6,000-7,000 words is a respectable vocabulary size, however this does not mean that all these words have been learnt through English. It is highly likely that a reasonable proportion of these words are cognates in English and Spanish and thus can be answered correctly on the test through knowledge of Spanish. Nonetheless, these words are unlikely to pose reading problems for Spanish speakers. (See Chapter 7 for further discussion).

2. *Lexical coverage of medical texts.* One of the most important findings of this thesis was that in order to achieve the 98% coverage of medical texts, an enormous number of medical words, i.e., between 26,000 and 27,000 medical words, beyond the 3,000 most frequent word families of the English language, are required for a good lexical comprehension of written medical textbooks. Because a large proportion of the running words in a medical text is made up of medical words, reading medical textbooks will be a challenging task requiring a variety of reading and vocabulary coping strategies.
3. *Medical word lists.* A total of twenty six 1,000 medical word lists – divided up into three 1,000 MEDGEN word lists and twenty three 1,000 MED word lists in Chapter 4 (see Appendix 2.1) – and twenty seven 1,000 medical lists – divided up into four 1,000 MEDGEN word lists and twenty three 1,000 MED word lists in Chapter 5 – were created and run through Range (Heatley et al., 2002) adding also some existing word lists (such as GSL, AWL, Pilot Science List, and BNC/COCA lists) to estimate the vocabulary load of medical texts written in English. A 98% lexical coverage of medical texts was reached after running both the twelve 1,000 most frequent medical word lists and the first five 1,000 BNC/COCA lists over the medical corpus (i.e., the med1 corpus) used to create these medical lists. A possibly adequate lexical coverage of 96% was achieved after running at least the same lists on an independent medical corpus (i.e., the med2 corpus). The good coverage results of the medical word lists over two different medical corpora indicate that the medical word lists created as part of the present study can be confidently used to obtain reliable estimates of the lexical demands of medical texts.
4. *Number of medical words.* Medical texts have a large topic-related vocabulary of over 32,195 medical word types (see Appendix 3.5). This represents a formidable learning goal for students of medicine both native speakers and non-native speakers. These medical words are spread across the high, mid and low-frequency levels, with

a large number not occurring in the most frequent 20,000 words of English. Students of medicine have a lot of words to learn as they study the demanding subject of medicine.

Let us know look at an extract from *Merck Manual of Diagnosis and Therapy* , 19th edition (Porter & Kaplan, 2011, Chapter 90) and identify the spread of words in this medical text across the three frequency levels with the high-frequency words (the most frequent 3,000 BNC/COCA word family lists) in **bold**, the mid-frequency words (from the fourth to the ninth 1,000 BNC/COCA lists) underlined, the low-frequency words (from the tenth to the twenty fifth 1,000 BNC/COCA lists) in *italics*, and the words off these twenty five 1,000 BNC/COCA lists unmarked. (See Figure 8.1).

Skin cancer is the most common type of cancer and usually develops in sun-exposed areas of skin. **The** incidence **is highest among outdoor workers**, sportsmen, **and** sunbathers **and is** inversely **related to the amount of** *melanin* **skin** pigmentation; **fair-skinned people are most** susceptible. **Skin cancers may also develop years after** therapeutic x-**rays or exposure to** carcinogens (eg, arsenic ingestion). **Over one million new cases of skin cancer are** diagnosed **in the** US **yearly**. **About** 80% **are** basal **cell** carcinoma, 16% **are** *squamous* **cell** carcinoma, **and** 4% **are** *melanoma*. Paget's **disease of the** nipple **or** extramammary Paget's (usually **near the** *anus*), Kaposi's *sarcoma*, tumors **of** adnexa, **and** *cutaneous* T-**cell** lymphoma (mycosis fungoides) **make up the remaining**, **less common**, **forms of skin cancer**. **Initially**, **skin cancers are often** asymptomatic. **The most frequent presentation is a** papule **or blind** pimple **that does not go away**. **Any** lesion **that appears to be** enlarging **should be** biopsied **whether tenderness**, **mild** inflammation, crusting, **or occasional bleeding is present or not**. **If treated early**, **most skin cancers are curable**. **Screening**: **routine screening for skin cancer is by patient self-examination**, physician **examination**, **or both**. **Prevention**: **because many skin cancers seem to be related to** ultraviolet (UV) **exposure**, **a number of measures are recommended to limit exposure**. **Sun avoidance**: **seeking shade**, minimizing **outdoor activities between** 10am **and** 4pm (when **sun**'s **rays are strongest**), **and avoiding** sunbathing **and the use of** tanning **beds**. **Use of protective clothing**: **long**-sleeved **shirt**, pants, **and broad**-brimmed **hat**. **Use of** sunscreen: **at least sun protection factor** (SPF) 30 **with** UVA **protection**, **used as directed**; **should not be used to** prolong **sun exposure**. **Current evidence is inadequate to determine whether these measures reduce** incidence **or mortality of** *melanoma*; **in** nonmelanoma **skin cancers** (basal **cell and** *squamous* **cell** carcinoma), **sun protection does decrease the** incidence **of new cancers**.

Figure 8.1 An extract from a medical textbook showing the spread of medical words across the high, mid and low-frequency levels (Porter & Kaplan, 2011, Chapter 90)

5. *The nature of specialised texts.* The vocabulary load of a specialised text is much lighter than the vocabulary load of a collection of texts on a diverse range of topics. Different topics involve different vocabulary, and keeping within a specialised field can reduce the number of different words dramatically, by around 50%. There is thus value in having special purposes courses for learners studying in the same subject area. Such courses avoid vocabulary that is not of immediate relevance to the subject being studied.

The methodology of this study demonstrates an efficient way of identifying technical vocabulary in specialised texts.

8.2 What should go into the vocabulary component of an English for Medical Purposes course for medical students who are not native speakers of English?

Vocabulary expansion should be an important goal for teachers of English for Medical Purposes. In order to help English for Medical Purposes learners better cope with the lexical demands of medical texts, as shown by the results of the present study, English for Medical Purposes teachers need to perform various roles in the language classroom. According to Nation (2008, p. 1), the vocabulary teacher's jobs in order of importance are as follows: planning, strategy training, testing and teaching vocabulary.

8.2.1 Planning the vocabulary component of an English for Medical Purposes reading course

According to Nation (2008, 2013), planning is the first and most important role of the L2 teacher. Planning is crucial when designing the lexical component of a language course, and involves deciding what needs to be learnt and how it is best learnt.

Planning should be easier when the teachers and the students share an L1. There is a great advantage in L2 teaching if the learners share the same L1 and if the L2 teacher speaks the L1 of the learners. Shared knowledge of the L1 of the learners should enable the teacher to use a direct L1 translation to provide the meanings of L2 words, to test the knowledge of L2 words, and also to judge how difficult particular words may be for the learners.

It is useful for English for Medical Purposes teachers to gather information on the learners' general vocabulary size (e.g., using the bilingual or monolingual VST), the types of words occurring in medical texts, and the learning burden of the words chosen to be focused on in the English for Medical Purposes classroom. This study suggests that the best time to move to a specialised medical focus is when learners know an average of 6,000 word families. In the present study, this figure roughly corresponds to what the learners about to enter medical study already know as measured by the Vocabulary Size Test. It would seem in the present circumstances that no change would be recommended to the entry requirements for English language for learners wishing to do medical study.

For learners who have a special purpose for learning English, a specialist focus reduces the amount of vocabulary to learn by avoiding general purpose vocabulary that is not relevant to the specialised subject area. In Chapter 4, we saw that a specialised medical corpus contains fewer than half of the general purpose words that occur in a varied corpus of similar length. It is thus efficient to move to a specialised focus as soon as enough general purpose vocabulary is known. This move to a specialised focus involves a wider consideration of the needs of the students. Moving to a specialised focus excludes other focuses, such as casual conversation, reading for pleasure, and pursuing hobbies through the medium of English. Both teachers and students need to consider what their needs really are.

The medical terms found in this study were accompanied by their frequency of occurrence in the medical corpus. As in most frequency-based lists, there is a wide range of frequencies roughly corresponding to Zipf's law (Zipf, 1949). That means that a very large proportion of the different words, roughly half according to Zipf's law, occur only once. It also means that there is a relatively small number of words that are very frequent. These high-frequency technical words give the greatest return for learning because the effort to learn them is repaid by the many opportunities to meet and use them.

However, the learning of technical terms is not easily done in isolation, because in order to understand the terms, we need to understand the field of knowledge that they represent. This means that English teachers who are providing courses to support the study of medicine should largely focus their attention on developing the learners' skill in applying strategies to deal with the vocabulary they meet during their study. English teachers are not

likely to have a depth of understanding of medicine that would allow them to successfully explain medical terms let alone deal with medical texts where these terms occur.

8.2.1.1 The four strands of a well-balanced English for Medical Purposes reading course

The vocabulary component of a language course is only one part of that course. When planning a well-balanced course, there is value in considering the four strands of a language course, namely, meaning-focused input, meaning-focused output, language-focused learning, and fluency development (Nation, 2007). Likewise, when planning the lexical component of an English for Medical Purposes course these four strands can be considered. In this respect, Nation (2013, p. 204) suggests that "when looking at the learning of a particular skill like reading, it is useful to look at what can make up such a course across the four strands."

Table 8.1 below includes some useful activities suggested by Nation (2008, 2013) for teaching vocabulary and reading, and that are also worth considering when planning an English for Medical Purposes reading course. In such a course, much of the time spent on meaning-focused output in a general course would be largely spent on meaning-focused input.

Table 8.1 Suggested activities in an English for Medical Purposes reading course for each of the four strands

Four strands	**Examples of activities that can be used in an English for Medical Purposes reading course**
Meaning-focused input focuses the learner's attention on the message of the reading materials	Extensive reading of medical texts Reading adapted medical texts Paired and group reading
Meaning-focused output focuses the learner's attention on discussing the messages of the written texts with others	Exchange of information on what was read or written Prepared talks Writing tasks
Language-focused learning focuses the learner's attention on the linguistic features of the written texts	Intensive reading of medical texts Deliberate study of reading strategies and vocabulary strategies Deliberate learning of vocabulary for reading medical texts Using word cards
Fluency development focuses the learner's attention on developing reading fluency	Speed reading Repeated reading Linked skills activities

Nation (2013) repeatedly emphasises the importance of maintaining a balance between the four strands of a course when looking at language learning in general, the learning of a specific language skill like reading, or the learning of specialised vocabulary from a particular subject field like medicine. In relation to the reading skill and the four strands, Nation (2013) recommends that 50% of the time in a reading course should be spent on meaning-focused input activities; that is, English for Medical Purposes learners should spend about half of the time paying attention to the message of their English for Medical Purposes readings. For this purpose, the materials should be at the right linguistic level of the learners and include few new topic-related words that can be understood or guessed using their background knowledge and some context clues. For example, for an introductory English for Medical Purposes reading course this could be achieved by adapting the medical texts used in class to the lexical knowledge and average vocabulary size of the group of L2 learners. Where adaptation is not feasible, extensive reading should focus on texts covering material where the content is familiar or at least partly known.

Meaning-focused output activities can be promoted by encouraging English for Medical Purposes students to have group discussions and to prepare talks on the medical topics that they have been reading about. Learning activities in the language-focused output strand may include the use of word cards or mnemonic techniques to deal with the learning of new medical words, and should certainly involve intensive reading of medical texts.

A total of 25% of the time in a course should be spent on fluency development. A way to promote fluency development is by engaging learners in linked skill activities. This type of activity provides opportunities for repeated encounters, through reading, writing, listening or speaking, with easy and familiar topic-related texts. The main goal of these repeated and linked skill activities is to encourage faster processing and faster retrieval of language items.

There is a variety of activities that could be specifically targeted in those classes where the language teachers are more inclined to use authentic and unsimplified reading materials. This range of activities includes narrow reading, elaboration, easification, negotiation, intensive reading, pre-teaching, vocabulary exercises and glossing. (For a detailed explanation of these activities see Nation, 2013, pp. 230–247).

8.2.2 Strategy training for an English for Medical Purposes reading course

Another role of the ESP teacher is strategy training. The most productive strategies are likely to be (1) making use of definitions which occur in the text (a large proportion of technical words are defined as they occur in the text or through their appearance in diagrams and pictures), (2) making use of prefixes, suffixes and stems to reduce the learning burden of technical terms, (3) drawing on known cognates between the L1 and English, (4) making use of deliberate learning strategies such as flash cards and flash card programs, and (5) making use of medical dictionaries not only to look up the meanings of words but to make these lookups memorable. Let us look briefly at each of these strategies.

Definitions in context. Chung and Nation (2004) found that around 3.1% of medical terms were defined in the text or appeared in labelled illustrations. Bramki and Williams (1984) and Flowerdew (1992) found a high proportion of definitions in technical discourse and also found a variety of definition forms. Teachers of English for special purposes could usefully focus on training learners to recognise and interpret these definitions. Some learners may require very little training in the strategy, while others struggle with it. A useful approach is to outline the classic definition format, namely, *A (technical term) is a (general category) which (defining features)*, and to compare the various forms of definition in a text with this format. An example of the classic definition format in a medical text is as follows: “Dermatitis *(technical term)* is superficial inflammation of the skin

(general category) characterized by redness, edema, oozing, crusting, scaling, and sometimes vesicles *(defining features)*." (Porter & Kaplan, 2011, Chapter 28).

Word parts. Many medical words make use of prefixes, suffixes and stems and some of these word parts will be very frequently used within the field. For example, prefixes like *mal-* (bad, ill) in *malabsorption, malformation, malnutrition*, and *hyper-* (above measure) in *hypertension*, *hypersensitivity*, *hypertropy*, stems like *neuro* (related to the nervous system) in *nonneuronal, antineural, aponeurosis*, and *nephro* (related to the kidneys) in *hypernephroma, hydronephrosis, subnephrotic,* or suffixes like *–itis* (inflammatory disease) in *arthritis*, *dermatitis*, *pancreatitis*, and *–oma* (tumour of abnormal growth) in *lymphoma*, *melanoma*, *sarcoma* commonly occur in medical words.

Learners should develop the attitude of looking at words analytically with the aim of seeing how the meanings of the parts contribute to the meaning of the whole. Frequent word parts should be deliberately learnt so that they are ready available for such analysis. English teachers can encourage this deliberate learning and provide practice in word part analysis.

Cognates. In the case of medical students who are L1 speakers of a Romance language like French, Italian, Portuguese, Spanish or Romanian, the learning burden of academic words and medical words is going to be lighter than for speakers of Asian languages like Japanese or Chinese. This is due to the large number of cognates coming from Latin and Greek that are shared between English and the Romance languages. Some examples of English/Romance language cognates in medical English are words like:

- The noun phrase '*a patient'* in English, *un patient* in French, *un pacient* in Catalan and Romanian, *um paciente* in Portuguese, *un paziente* in Italian comes from the Latin word *patiens*.
- The noun *encephalopathy* in English, *encéphalopathie* in French, *encefalopatia* in Catalan, Italian and Portuguese, *encefalopatía* in Spanish comes from the Greek words *enkephalos* (brain) + *pathos* (disease).

An example of a medical word with a heavier learning burden for speakers of Romance languages is the English/Romance language English non-cognate word '*marrow*', which is from Germanic origin. *Marrow* is the equivalent of *moelle* in French, *midollo* in Italian,

medul in Catalan, *medula* in Portuguese, and m*édula* in Spanish, which in these Romance languages comes from the Latin word *medulla*. Additionally, the word *medulla* is used in medical English in some particular contexts like *medulla oblongata,* and *renal medulla,* among others.

Speakers of Romance languages are likely to know a great deal of medical words in their L1 before they actually start learning them in English. This happens because a huge number of these content area (medical) words are initially the same in form and meaning across Romance languages. Nevertheless, cognates cannot always be treated in any language as if they were exact equivalent translations. Doing this may only hide the differences.

In Table 8.2 we suggest applying three criteria (i.e., frequency, usefulness, and familiarity) to each one of Nation's (2013) three aspects of knowing a word (i.e., form, meaning and use). This can be done to decide on the learning burden that more distant cognates, false cognates or non-cognates may pose on different L1 groups of English for Medical Purposes learners. Table 8.2 below summarises some of the questions language teachers could consider when looking at the learning burden of the medical vocabulary they are planning to emphasise with the English for Medical Purposes students.

Table 8.2 Three criteria (i.e., frequency, usefulness and familiarity) to estimate the learning burden of words. Adapted from aspects of knowing a word (Nation 2013, p.49)

Aspects of knowing a word	Criteria to estimate the learning burden	
Form	Frequency	How frequent is the word and its word parts?
	Usefulness	How useful is it to recognise the form of the word and its word parts?
	Familiarity	How familiar to the L1 of the learners is the form of the word and its word parts?
Meaning	Frequency	How frequent is the word found with the same meaning?
	Usefulness	How useful is it to know the meaning of the word?
	Familiarity	How close to the L1 of the learners is the meaning of the word?
Use	Frequency	How often is the word used?
	Usefulness	How useful for the message of the text is it to know how the word is used?
	Familiarity	How familiar to the L1 of the learners are the linguistic contexts in which the word is used?

Deliberate learning strategies. When large amounts of vocabulary need to be learnt within a rather short time, it is usually effective and efficient to make use of deliberate rote learning techniques such as the use of word cards or flash card programs. Although these deliberate learning activities are sometimes frowned on by teachers, the research evidence supporting their use is very strong (Nation, 2013, Chapter 11). Deliberate learning using word cards requires a small amount of training particularly in choosing what words to go on to the cards or into the programs, and in using spaced retrieval rather than massed learning when using the cards.

Using medical dictionaries. Because of the high quality dictionaries that are now produced, both bilingual and monolingual dictionaries represent reasonable choices for learners of English as a foreign language. Learners beginning medical study with a vocabulary size of 6,000 word families are likely to be capable of using monolingual English medical dictionaries. The main use of a dictionary in reading is to look up the meanings of words. However, dictionary look-up can also be used to establish the meanings of words through enriching the look-up by looking at all the given senses of the word to see its core meaning, by looking at the etymology of the word if it is given, by looking at example sentences containing the word and comparing these sentences with the context sentence that led to the look-up of the word. It is also useful to look at the immediate neighbours of the word in the dictionary to see if they share a similar meaning. There is an online medical dictionary at http://medical-dictionary.thefreedictionary.com, and online frequency level checkers like http://range.wordish.org, and http://language.tiu.ac.jp/flc that provide contextualised information on the various aspects of knowing a word.

Because of the very large number of technical words and the specialised knowledge required to understand medical English, teachers of English for Medical Purposes should aim to develop their learners' control of vocabulary learning strategies rather than concentrate on deliberately teaching medical vocabulary.

8.2.3 Testing the vocabulary component of an English for Medical Purposes reading course

In descending order of importance vocabulary testing should be the third focus of a teacher, after planning and strategy training. As previously discussed in Chapters 2 and 3, language teachers currently have some receptive vocabulary tests like the Vocabulary Levels Test (Nation, 1983), the Vocabulary Size Test (Nation & Beglar, 2007), and a series of Yes/No tests (Meara & Buxton, 1987; Meara & Jones, 1988, 1990; Meara & Milton, 2003; Meara & Miralpeix, 2006) that can be used either to determine the number of words L2 learners know (i.e., their vocabulary size) or the specific word list, and frequency level at which L2 learners need vocabulary training. If learners do not know the basic high-frequency vocabulary (the first 3,000 words of English), work on medical vocabulary will suffer particularly when dealing with vocabulary in medical texts.

In the context of the present study, the VST results (previously discussed in Chapter 3) show that the average vocabulary size of the 408 ESP learners was of around 6,000 words. Using this figure as one of the starting points for planning the vocabulary component of an English for Medical Purposes course, we can see that the immediate need is to focus on medical vocabulary. This vocabulary size means that a reasonable number of medical words (over 800 word families) may already be familiar to the learners because they occur in the first 6,000 words families of English.

It is worth monitoring learners' vocabulary growth both as a way of motivating their learning and seeing how close they are getting to having sufficient knowledge of medical terms.

8.2.4 Teaching the vocabulary component of an English for Medical Purposes reading course

The fourth role of the L2 teacher is actually teaching. Vocabulary teaching can have the following focuses:

1. *Teach useful word parts.* By conducting word part analysis when teaching vocabulary, language teachers bring the learners' attention to frequent prefixes,

suffixes and stems in medical texts. This gives learners the possibility of understanding a wider range of medical words using the same affixes, for example, by looking at common word parts and affixes in medical English.

2. *Teach high and mid-frequency words that occur in medical texts.* This implies the rearranging of general and academic word lists to include only the general and medical words occurring in medical texts. Once English for Medical Purposes students know the most frequent words in medical texts, the relisting of general and medical words based on their frequency and distribution, as well as the exclusion of general words that do not occur in medical texts, definitely increases the possibilities of focusing earlier on the enormous amount of medical words waiting to be learnt.
3. *Present the vocabulary in manageable formats.* Ideally, the focus on medical vocabulary should accompany the reading of medical texts. The frequent unknown words in the texts should then be the words focused on. Research on interference between items in lexical sets (Erten & Tekin, 2008; Tinkham, 1993, 1997; Waring, 1997) shows that bringing related words together makes learning more difficult. It is safer to focus on vocabulary as it occurs in texts than to bring closely related items together.
4. *Promote the use of a variety of reference sources.* Show students various written sources (e.g., specialised dictionaries, glossaries, concordances, topic-related lists of words, affixes, abbreviations or acronyms) or more informal oral sources (e.g. other teachers, specialists in the field, native speakers or other learners).
5. *Present medical words in relevant topic-related contexts.* Introduce the meaning and use of medical words in meaningful contexts.
6. *Draw attention to words within their typical phraseological patterns.* It is important to look at the role played by medical words considering the context and phraseological patterns (such as definition, classification, process, description, comparison/contrast, cause/effect, problem/solution, and ordering) commonly used to organise the information in expository medical texts. For example, definitions are frequently found in medical writing to introduce and explain the meaning of key technical terms. Nation (2013, p. 127) suggests that "teachers can help learners by clearly signalling the definitions they provide, by testing learners to diagnose how well they can recognise and interpret definitions, and by providing training in

recognising and interpreting definitions." Learners should be encouraged to identify and become familiar with the different ways how definitions are presented in medical texts.

7. *Promote extensive reading on medical topics.* Learners should be encouraged to do extensive reading on topics that include the vocabulary they are trying to learn.
8. *Promote repeated encounters with useful words.* Repetition is important in helping strengthen word knowledge. By rereading familiar texts with known words, or reading different texts on the same topic, or using online concordance tools like WebCorp Live (http://www.webcorp.org.uk/live/), learners have practical opportunities to have repeated instances of words and phrases that need to be reinforced in appropriate topic-related contexts.

Based on the large amount of medical vocabulary (at least 26,000 medical word types) required for good lexical comprehension, the teaching in the English for Medical Purposes classroom can only cover a small proportion of medical words that need to be learnt. For this reason, it is important that the English for Medical Purposes teacher sets ambitious vocabulary learning goals for the students.

It is essential for English for Medical Purposes teachers to promote the development of a variety of skills that equip medical students with the vocabulary learning strategies necessary to manage the acquisition of the enormous amount of words required to achieve good reading comprehension of medical textbooks.

8.3 Limitations of the study

The main limitations encountered during this research in relation to the administration of the Vocabulary Size Test, the compilation of the medical corpora, the capabilities of the computer software used for the lexical analysis, the identification of the medical words, and the creation of the medical word lists used can be summarised as follows:

1. *Administration of the Vocabulary Size Test.* The test used in this study was computer-based and this limited the number of students who could sit it, because students at higher education institutions with limited computer facilities and internet access could not sit the test. We also need to be a little cautious in interpreting the

results of the test because the multiple-choice recognition format of the test allows learners to gain higher scores than they would gain on a recall test which would require them to provide a translation for the tested words. In addition, we need to recognise that learners can answer questions correctly not because of previous learning of English but because of the cognates which exist in English and Spanish. It is important that these cognates are not excluded from the test, because they represent words that have a very low learning burden and thus can quickly become a part of the learners' receptive vocabulary. Learners of English who are native speakers of languages which are closely related to English, such as Spanish, French, Italian, Dutch and Swedish have a clear vocabulary advantage when learning English over speakers of other first languages.

2. *Capabilities of the computer software used.* The main limitation with the computer software (i.e., Range) used to obtain the lexical profile of the words in the medical corpus has to do with the impossibility of Range (Heatley et al., 2002) to be case sensitive and distinguish between upper case and lower case letter combinations. It hindered the possibility to identify an important number of abbreviations and acronyms (e.g., bid, ELISA, IgM, MRI, REM, tid, COPD, etc.) that are frequently used in medical texts. Moreover, it made the manual classification of the abbreviations, acronyms, and proper names in the medical corpus less reliable and more time-consuming. Additionally, it made problematic the identification of a number of abbreviations found throughout the most frequent word families in general English included in the existing general and academic word lists used. Another limitation of Range is that it does not distinguish homonyms or homographs but simply counts words on the basis of their form. A further limitation is that the Range software cannot count multiword units, and there are clearly multiword technical terms. From this viewpoint, the estimation of technical terms in medicine is evidently an underestimate.
3. *Identification of medical vocabulary.* It proved to be very demanding and time-consuming to distinguish medical words in the existing word lists and across the high, mid and low-frequency bands. The challenges of deciding where to draw the line between general and technical vocabulary led to the adoption, throughout this thesis, of various degrees of technicality and looked at the classification of medical vocabulary as part of a continuum. There is no doubt however that the technical

vocabulary of medicine is very large and represents a major learning burden for the student of medicine.

4. *Medical corpus limited to textbooks.* The medical texts included in the two medical corpora (i.e., the med1 and med2 corpora) compiled for the present investigation were restricted to textbooks.
5. *Validity of the bilingual English/Spanish Vocabulary Size Test.* An item analysis and comprehensive validation of the bilingual English/Spanish Vocabulary Size Test were not conducted. Further research is needed on these two aspects of the vocabulary size study.

8.4 **Further research**

As with any piece of research, the results raise as many questions as they answer, and further research would be valuable in dealing with these questions.

1. *Creation of specialised corpora for various subject-specific fields.* In the present study, we checked our results by developing an independent corpus of medical textbooks (i.e., the med2 corpus) to make sure that the results were reliable and generalisable. Even though, as previously explained in the review of the literature of the present investigation, some vocabulary researchers have already looked at the lexis of particular specialised subject-areas, there is still more research needed that investigates the frequency and distribution of general, academic, and topic-related technical words across the high, mid, and low-frequency bands using larger corpora (of at least 1 million running words) than those commonly used for lexical analysis of such corpora. It is also important to have more research that estimates the number of words needed within particular subject fields to achieve the optimal 98% lexical threshold.
2. *Development of subject-specific word lists.* It would be important to continue creating word lists for specific subject areas, like the ones created here for medicine. Subject-specific word lists can provide more accurate estimates of the lexical demands that specialised texts pose for ESP readers. Moreover, these word lists have been useful for drawing ESP teachers and learners' attention to the general, academic, and technical words needed for adequate lexical knowledge. Despite the lexical shortcuts that can be undoubtedly offered by subject-specific word lists, the

amount of technical words needed for good reading comprehension represents a major vocabulary learning goal for L2 English for Medical Purposes students. It is likely that the field of medicine is one of a few extreme cases regarding the vocabulary load of technical texts. Further research in other fields would be useful informing ESP course designers, test and material developers and language teachers of the lexical demands that other fields of study pose for ESP students.

3. *Subject-specific vocabulary size tests.* It would be useful to construct vocabulary size tests for specific subject fields. The development of these tests would provide more accurate estimates of the number of words ESP students actually know in a particular content area. Such developments would depend heavily on the type of research done on this thesis, namely, the creation of word lists to use for sampling for subject-specific vocabulary tests.
4. *Affixation and stems in medical words.* It would be worth investigating the most common prefixes, suffixes and stems occurring in medical words. We suspect that affixation of high-frequency medical words could well be very different to the most frequent affixes already identified for general English words. Some of this research has already been done (Collins & DePetris, 2013; Ehrlich & Schroeder, 2013; Hutton, 2006; Lankamp, 1988) but it needs to be done on a specific corpus in order to obtain frequency figures to guide the learning of the word parts.
5. *Cognate awareness.* The fact that some L1 Spanish learners of ESP consistently failed to identify some English/Spanish cognates, makes us believe that further research that investigates the level of students' morphological awareness of some high-frequency word parts found in cognates and non-cognates is still needed.
6. *Role of multiword units, and collocations*. The meaning of content area vocabulary is undoubtedly related to its lexical environment. Depending on the subject field in which technical words are used, they seem to collocate in very distinctive ways. When deciding on the technical nature of vocabulary, it may be crucial to look at the multiword units and linguistic contexts in which technical words typically occur. Thus, it would be useful for future research to identify the most common collocational patterns both in medicine and other subject-specific fields.
7. *Role of the English for Medical Purposes learners in the acquisition of the large number medical words.* Throughout this thesis and in the present chapter, the pedagogical implications of managing the enormous amount of vocabulary required

to read medical texts has been discussed primarily from the point of view of the English for Medical Purposes teacher. Nevertheless, it would also be useful to explore further the role that the English for Medical Purposes learners play in the acquisition of the vocabulary strategies required for achieving good lexical comprehension of medical texts written in English. Undoubtedly, it is essential for English for Medical Purposes leaners to embark on the independent learning and use of the vocabulary strategies needed inside and outside the classroom for being able to cope with the lexical demands of medical texts.

8.5 Closing remarks

The most striking finding of this thesis is that medical texts include a very large proportion (37%) of specialised words – more than one in every three running words is a medical word. As well as making up a large proportion of the text, there is a very large number of medical technical words. The study of medicine involves a very substantial vocabulary component. From a pedagogical perspective, the present research has looked at two aspects of needs analysis – lacks (vocabulary size of a group of ESP undergraduates), and necessities (lexical demands of medical texts). The corpus comparison approach to the investigation of the vocabulary of medical textbooks was adopted to be able to make more informed decisions on the selection criteria of the lexis (general, academic, and technical) that should be considered in the planning and implementation of the lexical syllabus of a reading course for learners of English for Medical Purposes.

For medical students to be able to cope with the lexical demands of medical texts and become autonomous readers in their field of specialisation, it is important that the ESP teacher prioritises strategy training in the ESP classroom, and teaches, with a few relevant specialised words, the vocabulary learning techniques required by the ESP learners for them to be able to continue to practice and learn independently the specific linguistic features in general and lexis in particular from their specialised content areas. This autonomous L2 practice/learning will enable ESP students in the end to be up to date during their university studies and future professional careers, and become proficient readers of their ESP occupational dialect.

The development of medical knowledge is a gradual process over a substantial period of time. Similarly the growth of medical vocabulary also involves a long term learning process that can be supported by training in dealing with technical vocabulary.

The present research is unique in developing a comprehensive twofold methodology (i.e., semantic rating scale and corpus comparison approach) which enabled the identification of a large number of words required to reach the 98% optimal lexical threshold of medical textbooks. With respect to the instructional value of the medical word lists created as part of this investigation on the vocabulary load of medical texts, these lists provide an opportunity for the construction of comprehensive measures for testing medical vocabulary. These medical word lists are a useful source of words that have been selected using well-informed word-list-creation principles, and can be confidently used when planning vocabulary strategy training of ESP reading classes for medical students. Likewise, the series of newly created medical word lists can be confidently used for enhanced curriculum design, materials development, and vocabulary testing in teaching learners of English for Medical Purposes.

Reference list

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Anderson, R. C., & Freebody, P. (1982). Reading comprehension and the assessment and acquisition of word knowledge. Technical Report No. 249. *Advances in Reading/language Research*, 1–51.

Anderson, R. C., & Nagy, W. E. (1992). The vocabulary conundrum. *American Educator: The Professional Journal of the American Federation of Teachers, 16*(4), 14–18.

Ard, J., & Homburg, T. (1992). Verification of language transfer. In S. M. Gass & L. Selinker (Eds.), *Language transfer in language learning* (Vol. 5, pp. 47–70). Amsterdam: John Benjamins Publishing Company.

Baker, M. (1988). Sub-technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language, 4*(2), 91–105.

Basturkmen, H. (2006). *Ideas and options in English for Specific Purposes*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Bauer, L. (1993). *The Wellington corpus of written New Zealand English (WWC)*. Wellington, New Zealand: Department of Linguistics, Victoria University of Wellington.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279. http://doi.org/10.1093/ijl/6.4.253

Becher, T. (1987). Disciplinary discourse. *Studies in Higher Education*, *12*(3), 261–274. http://doi.org/10.1080/03075078712331378052

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Velde, H. V. de. (2001). Examining the Yes/No vocabulary test: Some methodological issues in theory and practice. *Language Testing*, *18*(3), 235–274. http://doi.org/10.1177/026553220101800301

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101–118. http://doi.org/10.1177/0265532209340194

Bramki, D., & Williams, R. (1984). Lexical familiarization in economics text, and its pedagogic implications in reading comprehension. *Reading in a Foreign Language*, *2*(1), 169–181.

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

Brinton, D., Snow, M. A., & Wesche, M. B. (2003). *Content-based second language instruction*. Michigan: University of Michigan Press.

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

Campion, M. E., & Elley, W. B. (1971). *An academic vocabulary list*. Wellington: NZCER.

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research*, *10*(2), 157–187. http://doi.org/10.1177/026765839401000203

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. (2003). A corpus comparison approach for terminology extraction. *Terminology*, *9*(2), 221–245.

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.

Cobb, T. M. (2000). One size fits all? Francophone learners and English vocabulary tests. *Canadian Modern Language Review/ La Revue Canadienne Des Langues Vivantes*, *57*(2), 295–324. http://doi.org/10.3138/cmlr.57.2.295

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

Collins, C. E., & DePetris, A. (2013). *A short course in medical terminology* (Third edition). Philadelphia, PA: Lippincott Williams & Wilkins.

Cowan, J. R. (1974). Lexical and syntactic research for the design of EFL reading materials. *TESOL Quarterly*, *8*(4), 389–399. http://doi.org/10.2307/3585470

Coxhead, A. (1998). *An academic word list* (Vol. 18). Wellington, New Zealand: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

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

Coxhead, A. (2011). The Academic Word List 10 years on: Research and teaching implications. *TESOL Quarterly*, *45*(2), 355–362. http://doi.org/10.5054/tq.2011.254528

Coxhead, A. (2012). Researching vocabulary in secondary school English texts: The hunger games and more. *English in Aotearoa*, *October*, 34–41.

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

Coxhead, A., Stevens, L., & Tinkle, J. (2010). Why might secondary science textbooks be difficult to read? *New Zealand Studies in Applied Linguistics*, *16*(2), 37–52.

Csilla, K. (2009). English as the lingua franca of medicine. In C. Tanulmányok, H. Szaknyelvoktatásról, & E. Kutatásról (Eds.), *Porta Lingua - 2009* (pp. 53–64). Debrecen: Central Print Nyomda Debrecen.

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *Special Issue : ESP in Asia*, *33*(0), 66–76. http://doi.org/10.1016/j.esp.2013.08.001

Davies, M. (2014). *The Corpus of Contemporary American English: 450 million words (1990-2014)*. Retrieved from http://corpus.byu.edu/coca/

Douglas, S. R. (2014). After the first 2,000: A response to Horst's "Mainstreaming second language vocabulary acquisition." *Canadian Journal of Applied Linguistics / Revue Canadienne de Linguistique Appliquée*, *16*(1), 189–199.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, *28*(3), 157–169. http://doi.org/10.1016/j.esp.2009.02.002

Ehrlich, A., & Schroeder, C. L. (2013). *Medical terminology for health professions* (Seventh edition). Boston, MA: Cengage Learning.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*(2), 253–272. http://doi.org/10.1177/0265532212459028

Elgort, I., & Coxhead, A. (in press). An introduction to the Vocabulary Size Test: Description, application and evaluation. In J. Fox & V. Aryadoust (Eds.), *Current trends in language testing in the Pacific Rim and the Middle East: Policies, analysis, and diagnosis*. Cambridge: Cambridge Scholars Publishing.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

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

Erten, I. H., & Tekin, M. (2008). Effects on vocabulary acquisition of presenting new words in semantic sets versus semantically unrelated sets. *System: An International Journal of Educational Technology and Applied Linguistics*, *36*(3), 407–422.

Eyckmans, J. (2004). *Measuring receptive vocabulary size reliability and validity of the yes/no vocabulary test for French-speaking learners of Dutch* (Unpublished doctoral dissertation). Volumen 19 of LOT Publications. Katholieke Universiteit Nijmegen, The Netherlands. Retrieved from http://www.lotpublications.nl/index3.html

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

Flowerdew, J. (1992). Definitions in science lectures. *Applied Linguistics, 13*(2), 202–221. http://doi.org/10.1093/applin/13.2.202

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*, 127–143.

Fraser, S. (2009). Breaking down the divisions between general, academic, and technical vocabulary: The establishment of a single, discipline-based word list for ESP learners. *Hiroshima Studies in Language and Language Education, 12*, 151–167.

Fraser, S. (2013). Building corpora and compiling pedagogical lists for university medical students. *Hiroshima Studies in Language and Language Education, 16*, 65–88.

Fraser, S., Davies, W., & Tatsukawa, K. (2015). Medical word list development through corpus and course construction. *Hiroshima Studies in Language and Language Education*, *18*, 179–193.

Frînculescu, I. C. (2009). The physiology of English as a lingua franca in medicine. *Fiziologia-Physiology*, *19*(2), 4–6.

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

Ghadessy, M. (1979). Frequency counts, word lists, and materials preparation: a new approach. *English Teaching Forum*, *17*(1), 24–27.

Ghalandari, S., & Talebinejad, M. R. (2012). Medical ESP textbook evaluation in Shiraz medical college. *Education Research Journal*, *2*(1), 20–29.

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/

Graves, M. F., & Watts-Taffe, S. (2008). For the love of words: Fostering word consciousness in young readers. *The Reading Teacher*, *62*(3), 185–193.

Graves, M. F., & Watts-Taffe, S. M. (2002). The place of word consciousness in a research-based vocabulary program. In A. E. Farstrup & S. J. Samuels (Eds.),

What research has to say about reading instruction (pp. 140–165). Newark, DE: International Reading Association.

Gu, P. Y. (2005). *Vocabulary learning strategies in the Chinese EFL context*. Singapore: Marshall Cavendish Academic.

Gyllstad, H., Vilkaite, L., & Schmitt, N. (in press). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL International Journal of Applied Linguistics*, *166*(2).

Haliza, E., Ibrahim, E., Othman, K., Sarudin, I., & Muhamad, A. J. (2013). Measuring the vocabulary size of Muslim pre-university students. *World Applied Sciences Journal*, *1*, 44–49. http://doi.org/10.5829/idosi.wasj.2013.21.sltl.2136

Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System*, *37*(4), 614–626. http://doi.org/10.1016/j.system.2009.09.006

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.

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://doi.org/10.1177/1362168813494121

Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *Special Issue: ESP in Asia*, *33*(0), 54–65. http://doi.org/10.1016/j.esp.2013.07.001

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

Hutchinson, T., & Waters, A. (1987). *English for specific purposes*. Cambridge: Cambridge University Press. Retrieved from https://books.google.co.nz/books?id=7OvEeiyxNgEC

Hutton, A. R. (2006). *An introduction to medical terminology for health care: A self-teaching package.* (Fourth edition). Edinburgh: Churchill Livingstone.

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. (2002). Specificity revisited: how far should we go now? *English for Specific Purposes*, *21*(4), 385–395. http://doi.org/10.1016/S0889-4906(01)00028-X

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

Hyland, K. (2013). Writing in the university: education, knowledge and reputation. *Language Teaching*, *46*(01), 53–70.

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

Javid, C. Z. (2011). EMP needs of medical undergraduates in a Saudi context. *Kashmir Journal of Language Research, 14*(1), 89–110.

Karami, H. (2012). The development and validation of a bilingual version of the Vocabulary Size Test. *RELC Journal, 43*(1), 53–67. http://doi.org/10.1177/0033688212439359

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

Konstantakis, N. (2010). *Constructing a word list for the academic domain of business* (Unpublished doctoral dissertation). University of Wales, Swansea, UK.

Kucera, H., & Francis, W. N. (1964). *A standard corpus of present-day edited American English, for use with digital computers. (Brown corpus)*. Providence, Rhode Island: Cooperative Research Program of the U.S. Office of Education & Brown University.

Lankamp, R. E. (1988). *A Study on the effect of terminology on L2 Reading Comprehension: Should specialist terms in medical texts be avoided?* Amsterdam: Rodopi B.V.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? *Special Language: From Humans Thinking to Thinking Machines*, 316–323.

Laufer, B. (1990). Why are some words more difficult than others? Some intralexical factors that affect the learning of words. *IRAL - International Review of Applied Linguistics in Language Teaching*, *28*(4), 293–308. http://doi.org/10.1515/iral.1990.28.4.293

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. (1997). What's in a word that makes it hard or easy? Intralexical factors affecting the difficulty of vocabulary acquisition. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 140–155). Cambridge: Cambridge University Press.

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.

Leech, G., Johansson, S., & Hofland, K. (1978). *The Lancaster-Oslo/Bergen corpus of British English for use with digital computers. (LOB corpus)*. Lancaster University, University of Oslo, University of Bergen: Longman Group Limited, the British Academy, Department of British and American Studies, University of Oslo, Norwegian Research Council for Science and the Humanities, Norwegian Computing Centre for the Humanities.

Li, L., & MacGregor, L. J. (2010). Investigating the receptive vocabulary size of university-level Chinese learners of English: How suitable is the Vocabulary

Levels Test? *Language and Education*, *24*(3), 239–249. http://doi.org/10.1080/09500781003642478

Li, Y., & Qian, D. D. (2010). Profiling the Academic Word List (AWL) in a financial corpus. *System*, *38*(3), 402–411. http://doi.org/10.1016/j.system.2010.06.015

Lynn, R. W. (1973). Preparing word-lists: A suggested method. *RELC Journal*, *4*(1), 25–28. http://doi.org/doi:10.1177/003368827300400103

Maher, J. (1986). The Development of English as an international language of medicine. *Applied Linguistics*, *7*(2), 206–218. http://doi.org/10.1093/applin/7.2.206

Mair, C., & Ludwigs, A. (1999a). *The Freiburg-Brown corpus (Frown corpus)*. Freiburg: Universität Freiburg.

Mair, C., & Ludwigs, A. (1999b). *The Freiburg-LOB corpus (FLOB corpus)*. Freiburg: Universität Freiburg.

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. http://doi.org/10.1016/j.esp.2009.04.003

Meara, P. M., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142–154. http://doi.org/10.1177/026553228700400202

Meara, P. M., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied Linguistics in Society* (pp. 80–87). London: Centre for Information on Language Teaching and Research.

Meara, P. M., & Jones, G. (1990). *Eurocentres vocabulary size test 10KA*. Zurich: Eurocentres Learning Service.

Meara, P. M., & Milton, J. L. (2003). *X_Lex: The Swansea levels test*. Newbury: Express.

Meara, P. M., & Miralpeix, I. (2006). *Y_Lex: The Swansea advanced Vocabulary Levels Test. v2.05*. en, Swansea: Lognostics.

Miller, R. (2000). Beyond reductionism: The emerging holistic paradigm in education. *The Humanistic Psychologist*, *28*(1-3), 382–393. http://doi.org/10.1080/08873267.2000.9977003

Millward, C. M., & Hayes, M. (2012). *A biography of the English language*. Boston, MA: Cengage Learning.

Milton, J. L., & Treffers-Daller, J. (2013). Vocabulary size revisited: The link between vocabulary size and academic achievement. *Applied Linguistics Review*, *4*(1), 151–172.

Minshall, D. E. (2013). *A computer science word list* (Unpublished master's thesis). Swansea University, Swansea, UK. Retrieved from http://www.baleap.org/media/uploads/dissertation-awards/2014/Daniel_Minshall_Dissertation_Final_Draft.pdf

Miralpeix, I., & Meara, P. M. (2013). Knowledge of written form. In J. L. Milton & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge* (pp. 33–47). London: Palgrave Macmillan.

Mochida, K., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*(1), 73–98. http://doi.org/10.1191/0265532206lt321oa

Morgan, B. Q., & Oberdeck, L. M. (1930). Active and passive vocabulary. In E. W. Bagster-Collins (Ed.), *Studies in modern language teaching: Reports prepared for the modern foreign language study and the Canadian committee on modern languages* (Vol. 16, pp. 213–221). New York: Macmillan.

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

Murphy, J. M., & Kandil, M. (2004). Word-level stress patterns in the academic word list. *System*, *32*(1), 61–74. http://doi.org/10.1016/j.system.2003.06.001

Nacera, A. (2010). Languages learning strategies and the vocabulary size. *Procedia - Social and Behavioral Sciences*, *2*(2), 4021–4025. http://doi.org/10.1016/j.sbspro.2010.03.634

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

Nation, I. S. P. (Ed.). (1984). *Vocabulary lists: words, affixes and stems*. Occasional publication. Number 12. Wellington, New Zealand: English Language Institute, Victoria University of Wellington.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston, MA: Heinle & Heinle Publishers.

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

Nation, I. S. P. (2005). *Information on the British National Corpus list 14,000*. Wellington, New Zealand: Victoria University of Wellington. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

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. (2007). The four strands. *Innovation in Language Learning and Teaching*, *1*(1), 2–13. http://doi.org/10.2167/illt039.0

Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques*. Boston, MA: Heinle, Cengage Learning.

Nation, I. S. P. (2012). *The Vocabulary Size Test*. (23 October 2012). Unpublished paper. Wellington, New Zealand: Victoria University of Wellington. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf

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

Nation, I. S. P., & Anthony, L. (2013). Mid-frequency readers. *Journal of Extensive Reading*, *1*. Retrieved from http://jalt-publications.org/access/index.php/JER/article/view/868/40

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–12.

Nation, I. S. P., & Coxhead, A. (2014). Vocabulary size research at Victoria University of Wellington, New Zealand. *Language Teaching*, *47*(03), 398–403. http://doi.org/10.1017/S0261444814000111

Nation, I. S. P., & Gu, P. Y. (2007). *Focus on vocabulary*. Sydney, Australia: National Centre for English Language Teaching and Research, Macquaire University.

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

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

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

Nemati, A. (2010). Proficiency and size of receptive vocabulary: Comparing EFL and ESL environments. *International Journal of Educational Research and Technology*, *1*(1), 46–53.

Nguyen, L. T. C., & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal*, *42*(1), 86–99. http://doi.org/10.1177/0033688210390264

Niazi, M. M. (2014). The Need for English language courses in Pakistani medical colleges. *New Horizons*, *8*(1), 39.

Nordin, N. R. M., Stapa, S. H., & Darus, S. (2012). Are my words good enough to eat?: The teaching and learning of specialized vocabulary in culinary studies. In *Procedia - Social and Behavioral Sciences.* (pp. 22–23). Paper presented at the International Conference of Social Sciences & Humanities (ICOSH). UKM Bangi, Malaysia: Universiti Kebangsaan Malaysia.

Paquot, M. (2007). Towards a productively-oriented academic word list. In J. Waliński, K. Kredens, & S. Goźdź-Roszkowski (Eds.), *Corpora and ICT in Language Studies. PALC 2005* (Vol. 13, pp. 127–140). Frankfurt am Main: Peter Lang.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, 0265532212438053. http://doi.org/10.1177/0265532212438053

Peters, P., Collins, P., & Blair, D. (1986). *The Australian corpus of English (ACE)*. Sydney, Australia: Macquarie University.

Popa, D. E. (2013). Medical discourse and ESP courses for Romanian nursing undergraduates. *Procedia-Social and Behavioral Sciences*, *83*, 17–24.

Porter, R. S., & Kaplan, J. L. (Eds.). (2011). *The Merck manual of diagnosis and therapy* (19th edition). Whitehouse Station, NJ: Merck Sharp & Dohme Corp. Retrieved from http://www.merckmanuals.com/professional/index.html

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

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal*, *19*(2), 12–25. http://doi.org/10.1177/003368828801900202

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.

Richards, J. C. (1974). Word Lists: Problems and prospects. *RELC Journal*, *5*(2), 69–84. http://doi.org/10.1177/003368827400500207

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*(1), 77. http://doi.org/10.2307/3585941

Rutherford, W. E., & Sharwood Smith, M. (1985). Consciousness-raising and universal grammar. *Applied Linguistics*, *6*(3), 274–282. http://doi.org/10.1093/applin/6.3.274

Salager-Meyer, F. (1983). The lexis of fundamental medical English: classificatory framework and rhetorical function (a statistical approach). *Reading in a Foreign Language*, *1*(1), 54–64.

Salager-Meyer, F. (1985). Specialist medical English lexis: Classificatory framework and rhetorical functions - a statistical approach. *EMP Newsletter*, *2*(2), 5–18.

Salager-Meyer, F. (2014). Origin and development of English for Medical Purposes. Part I: Research on written medical discourse. *Medical Writing*, *23*(1), 49–51. http://doi.org/10.1179/2047480613Z.000000000187

Salager-Meyer, F., Llopis de Segura, G., & Guerra Ramos, R. (in press). EAP in Latin America. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for Academic Purposes*. London: Routledge.

Schmidt, R. (1992). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, *13*, 206–226.

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

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

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88. http://doi.org/10.1177/026553220101800103

Scott, J. A., & Nagy, W. E. (2004). Developing word consciousness. In E. J. Kame'enui & J. F. Baumann (Eds.), *Vocabulary instruction: Research to practice* (pp. 201–217). New York: Guilford Press.

Sevigny, P., & Ramonda, K. (2012). Vocabulary: What Should We Test? In N. Sonda & A. Krause (Eds.), *JALT2012 conference proceedings* (pp. 701–711). Tokyo: JALT. Retrieved from http://jalt-publications.org/proceedings/articles/3320-vocabulary-what-should-we-test

Shastri, S. V. (1986). *The Kolhapur corpus of Indian English for use with digital computers (KCIE)*. Kolhapur, India: Department of English Shivaji University.

Shillaw, J. (1996). The application of Rasch modelling to yes/no vocabulary tests. Unpublished discussion paper. Vocabulary acquisition research group at Swansea University. Retrieved from http://www.lognostics.co.uk/vlibrary

Shillaw, J. (2009). Putting yes/no tests in context. In T. Fitzpatrick & A. Barfield (Eds.), *Lexical processing in second language learners: Papers and perspectives in honour of Paul Meara* (pp. 13–24). Bristol, UK: Multilingual Matters.

Showell, C., Cummings, E., & Turner, P. (2010). Language fames and patient-centred eHealth. *Studies in Health Technology and Informatics*, (155), 55–61.

Sonchaiya, P., Wasuntarasophit, S., & Chindaprasirt, A. (2011). Technical and academic vocabulary from food technology research articles: Challenges for EAP/ESP teachers. *Interdisciplinary Discourses in Language and Communication*, 322–334.

Sorell, J. (2012). Zipf's Law and vocabulary. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Oxford: Blackwell Publishing Ltd. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal1302/abstract

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*(3), 271–282. http://doi.org/10.1080/15434303.2014.922977

Stoddard, G. D. (1929). An experiment in verbal learning. *Journal of Educational Psychology*, *20*(6), 452–457. http://doi.org/10.1037/h0073293

Sutarsyah, C., Nation, I. S. P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based case study. *RELC Journal*, *25*(2), 34–50.

Tinkham, T. (1993). The effect of semantic clustering on the learning of second language vocabulary. *System*, *21*(3), 371–380. http://doi.org/10.1016/0346-251X(93)90027-E

Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research*, *13*(2), 138–163. http://doi.org/10.1191/026765897672376469

Vongpumivitch, V., Huang, J., & Chang, Y.-C. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, *28*(1), 33–41. http://doi.org/10.1016/j.esp.2008.08.003

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. http://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. (2007). Collocation and technicality in EAP engineering. *Journal of English for Academic Purposes*, *6*(1), 18–35. http://doi.org/10.1016/j.jeap.2006.10.001

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

Waring, R. (1997). A comparison of the receptive and productive vocabulary sizes of some second language learners. *Immaculata (Occasional Papers of Notre Dame Seishin University, Okayama), 1*, 53–68.

Warrell, D. A., Cox, T. M., Firth, J. D., & Ogg, G. S. (Eds.). (2010). *Oxford textbook of Medicine* (5th edition). Oxford: Oxford University Press. Retrieved from http://oxfordmedicine.com/view/10.1093/med/9780199204854.001.1/med-9780199204854

Webb, S. A., & Nation, I. S. P. (2008). Evaluating the vocabulary load of written texts. *TESOLANZ Journal*, *16*, 1–10.

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

Willis, J., & Willis, D. (1996). *Challenge and change in language teaching*. London: Heineman.

Woodward-Kron, R. (2008). More than just jargon – the nature and role of specialist language in learning disciplinary knowledge. *Journal of English for Academic Purposes*, *7*(4), 234–249. http://doi.org/10.1016/j.jeap.2008.10.004

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215–229.

Zhang, X. (2013). The I don't know option in the Vocabulary Size Test. *TESOL Quarterly*, *47*(4), 790–811. http://doi.org/10.1002/tesq.98

Zhang, X., & Lu, X. (2014). A longitudinal study of receptive vocabulary breadth knowledge growth and vocabulary fluency development. *Applied Linguistics*, *35*(3), 283–304. http://doi.org/10.1093/applin/amt014

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. New York: Hafner.

Appendices

Appendix 1: The Spanish bilingual version of the VST

Appendix 1.1 Translation of the 140-item localised for Spanish VST

Instrucciones = Instructions
Seleccione la opción en español (sinónimo, definición o traducción) que mejor describa el significado de la palabra solicitada.
= Choose the answer (synonym or definition) with the closest meaning to the key word in the question.

Evite dejar preguntas sin responder a lo largo de la prueba.
= While taking the test try to answer all the questions.

First 1000

1.	**see**: They saw it.	
	a.	cortaron
	b.	esperaron
✓	**c.**	**vieron**
	d.	empezaron

2.	**time**: They have a lot of time.	
	a.	dinero
	b.	comida
✓	**c.**	**horas**
	d.	amigos

3.	**period**: It was a difficult period.	
	a.	pregunta
✓	**b.**	**tiempo**
	c.	hecho
	d.	libro

4.	**figure**: Is this the right figure?	
	a.	respuesta
	b.	lugar
	c.	momento
✓	**d.**	**número**

5.	**poor**: We are poor.	
✓	**a.**	**no tenemos dinero**
	b.	nos sentimos felices
	c.	estamos muy interesados
	d.	no nos gusta trabajar mucho

6.	**drive**: He drives fast.	
	a.	nada
	b.	aprende
	c.	lanza la pelota
✓	**d.**	**conduce el automóvil**

7.	**jump**: She tried to jump.	
	a.	flotar en el agua
✓	**b.**	**elevarse del suelo con un impulso rápido**
	c.	dejar el automóvil aparcado al borde del camino
	d.	ir muy rápido

8.	**shoe**: Where is your shoe?	
	a.	la persona que cuida de usted
	b.	el objeto donde guarda el dinero
	c.	lo que usa para escribir
✓	**d.**	**lo que se pone en los pies**

9.	**standard**: Her standards are very high.	
	a.	pieza del zapato en la parte del talón
	b.	calificaciones que obtiene en la escuela
	c.	dinero que solicita
✓	**d.**	**niveles que alcanza en todo**

10.	**basis**: I don't understand the basis.	
	a.	presupuesto
	b.	palabras
	c.	señales de tránsito
✓	**d.**	**fundamento**

Second 1000

1.	**maintain**: Can they maintain it?	
✓	**a.**	**conservarlo como está**
	b.	agrandarlo
	c.	adquirir uno mejor
	d.	obtenerlo

2.	**stone**: He sat on a stone.	
✓	**a.**	**superficie dura**
	b.	tipo de silla
	c.	superficie blanda sobre el suelo
	d.	parte del árbol

3.	**upset**: I am upset.	
	a.	cansado
	b.	famoso
	c.	rico
✓	**d.**	**molesto**

4.	**drawer**: The drawer was empty.	
✓	**a.**	**cajón corredizo**
	b.	lugar para guardar los automóviles
	c.	aparato para mantener los productos fríos
	d.	recinto para animales

5.	**patience**: He has no patience.	
✓	**a.**	**no le gusta esperar**
	b.	no tiene tiempo libre
	c.	no tiene fe
	d.	no sabe lo que es justo

6.	**nil**: His mark for that question was nil.	
	a.	muy mala
✓	**b.**	**cero**
	c.	muy buena
	d.	regular

7.	**pub**: They went to the pub.	
✓	**a.**	**lugar para ir a beber y conversar**
	b.	institución en la que se guarda el dinero
	c.	edificio grande que tiene muchas tiendas
	d.	complejo destinado a la natación

8.	**circle**: Make a circle.	
	a.	boceto inicial
	b.	espacio en blanco
✓	**c.**	**forma redonda**
	d.	agujero grande

9.	**microphone**: Please use the microphone.	
	a.	aparato para calentar comida
✓	**b.**	**aparato para aumentar el sonido**
	c.	aparato que hace ver los objetos más grandes
	d.	aparato telefónico portátil

10.	**pro**: He's a pro.	
	a.	persona contratada para averiguar información secreta
	b.	persona estúpida
	c.	persona que escribe para un periódico
✓	**d.**	**persona a quien le pagan por practicar un deporte**

Third 1000

1.	**soldier**: He is a soldier.	
	a.	persona que tiene un negocio
	b.	persona que estudia
	c.	persona que trabaja con metal
✓	**d.**	**persona que está en el ejército**

2.	**restore**: It has been restored.	
	a.	dicho de nuevo
	b.	asignado a otra persona
	c.	rebajado de precio
✓	**d.**	**dejado como nuevo**

3.	**jug**: He was holding a jug.	
✓	**a.**	**recipiente para verter líquidos**
	b.	conversación informal
	c.	gorra de textura suave
	d.	arma de fuego

4.	**scrub**: He is scrubbing it.	
	a.	rayándolo
	b.	reparándolo
✓	**c.**	**restregándole con fuerza para limpiarlo**
	d.	haciéndole un dibujo simple

5.	**dinosaur**: The children were pretending to be dinosaurs.	
	a.	ladrones que roban en el mar
	b.	criaturas diminutas con alas y apariencia humana
	c.	criaturas grandes con alas, que escupen fuego
✓	**d.**	**animales que vivieron hace mucho tiempo atrás**

6.	**strap**: He broke the strap.	
	a.	promesa
	b.	cubierta superior
	c.	recipiente plano para servir los alimentos
✓	**d.**	**pieza larga y estrecha de tela para unir cosas**

7.	**pave**: It was paved.	
	a.	prohibido el paso
	b.	dividido
	c.	revestido con bordes dorados
✓	**d.**	**cubierto con una superficie dura**

8.	**dash**: They dashed over it.	
✓	**a.**	**fueron deprisa**
	b.	fueron despacio
	c.	pelearon
	d.	ojearon

9.	**rove**: He couldn't stop roving.	
	a.	emborracharse
✓	**b.**	**viajar**
	c.	cantar con los labios cerrados
	d.	trabajar mucho

10.	**lonesome**: He felt lonesome.	
	a.	ingrato
	b.	muy cansado
✓	**c.**	**solitario**
	d.	con mucha energía

Fourth 1000

1.	**compound**: They made a new compound.	
	a.	acuerdo
✓	**b.**	**mezcla de dos o más partes**
	c.	grupo de personas que forman un negocio
	d.	pronóstico basado en experiencias pasadas

2.	**latter**: I agree with the latter.	
	a.	hombre que dirige la iglesia
	b.	razón dada
✓	**c.**	**último**
	d.	respuesta

3.	**candid**: Please be candid.	
	a.	sea cuidadoso
	b.	sea compasivo
	c.	sea justo
✓	**d.**	**sea franco**

4.	**tummy**: Look at my tummy.	
	a.	trozo de tela para cubrir la cabeza
✓	**b.**	**estómago**
	c.	animal pequeño y peludo
	d.	pulgar

5.	**quiz**: We made a quiz.	
	a.	objeto para llevar las flechas
	b.	error grave
✓	**c.**	**serie de preguntas**
	d.	caja en donde los pájaros hacen su nido

6.	**input**: We need more input.	
✓	**a.**	**información, energía, etc. que se le suministra a algo**
	b.	trabajadores
	c.	relleno artificial para tapar un agujero en la madera
	d.	dinero

7.	**crab**: Do you like crabs?	
✓	**a.**	**criaturas marinas que caminan de lado**
	b.	galletas muy pequeñas y finas
	c.	cuellos duros y ajustados
	d.	insectos grandes de color negro que cantan por la noche

8.	**vocabulary**: You will need more vocabulary.	
✓	**a.**	**palabras**
	b.	destreza
	c.	dinero
	d.	armas

9.	**remedy**: We found a good remedy.	
✓	**a.**	**forma de solucionar el problema**
	b.	lugar público donde se sirven comidas
	c.	forma de preparar comida
	d.	regla matemática

10.	**allege**: They alleged it.	
✓	**a.**	**lo afirmaron sin tener pruebas**
	b.	le robaron las ideas a otra persona
	c.	presentaron hechos que lo demuestran
	d.	se expresaron en contra de los hechos que lo sustentaban

Fifth 1000

1.	**deficit**: The company had a large deficit.	
✓	**a.**	**gastó mucho más dinero del que ganó**
	b.	perdió mucho de su valor
	c.	tenía un plan de gastos para el cual usó mucho dinero
	d.	tenía mucho dinero guardado en el banco

2.	**weep**: He wept.	
	a.	se graduó
✓	**b.**	**lloró**
	c.	murió
	d.	se preocupó

3.	**nun**: We saw a nun.	
	a.	criatura larga y delgada que vive en la tierra
	b.	accidente terrible
✓	**c.**	**mujer con votos religiosos que lleva una vida estricta**
	d.	inexplicable luz brillante en el cielo

4.	**haunt**: The house is haunted.	
	a.	llena de adornos
	b.	alquilada
	c.	vacía
✓	**d.**	**llena de fantasmas**

5.	**compost**: We need some compost.	
	a.	buen apoyo
	b.	ayuda para sentirnos mejor
	c.	mezcla de piedras y arena que al fraguar adquiere dureza
✓	**d.**	**materia vegetal descompuesta**

6.	**cube**: I need one more cube.	
	a.	objeto puntiagudo empleado para unir cosas
✓	**b.**	**bloque sólido de caras cuadradas**
	c.	taza alta que viene sin plato
	d.	pedazo de cartón delgado doblado por la mitad

7.	**miniature**: It is a miniature.	
✓	**a.**	**algo muy pequeño en su clase**
	b.	instrumento para observar objetos diminutos
	c.	microorganismo vivo
	d.	línea corta para unir las letras en la escritura

8.	**peel**: Shall I peel it?	
	a.	dejarlo en agua por un tiempo
✓	**b.**	**pelarlo**
	c.	blanquearlo
	d.	cortarlo en pedazos finos

9.	**fracture**: They found a fracture.	
✓	**a.**	**ruptura**
	b.	pedacito
	c.	abrigo corto
	d.	joya extraña

10.	**bacterium**: They didn't find a single bacterium.	
✓	**a.**	**microorganismo que causa enfermedades**
	b.	planta de flores rojas y anaranjadas
	c.	animal que lleva agua en bultos sobre el lomo
	d.	artículo robado y luego vendido en una tienda

Sixth 1000

1.	**devious**: Your plans are devious.	
✓	**a.**	**engañosos**
	b.	bien desarrollados
	c.	no muy bien pensados
	d.	más costosos de lo necesario

2.	**premier**: The premier spoke for an hour.	
	a.	persona que trabaja en un tribunal
	b.	profesor universitario
	c.	aventurero
✓	**d.**	**mandatario**

3.	**butler**: They have a butler.	
✓	**a.**	**mayordomo**
	b.	máquina usada para talar árboles
	c.	profesor privado
	d.	pieza oscura y fría situada debajo de la casa

4.	**accessory**: They gave us some accessories.	
	a.	documentos que nos dan el derecho a entrar en un país
	b.	órdenes oficiales
	c.	ideas entre las que se puede escoger
✓	**d.**	**piezas adicionales**

5.	**threshold**: They raised the threshold.	
	a.	bandera
✓	**b.**	**punto o línea de inicio de algo**
	c.	parte interna del techo de una edificación
	d.	cantidad pagada por el uso de dinero recibido en préstamo

6.	**thesis**: She has completed her thesis.	
✓	**a.**	**extenso informe escrito para obtener un título universitario**
	b.	palabras pronunciadas por un juez al final de un juicio
	c.	primer año de empleo después de certificarse como profesor
	d.	tratamiento hospitalario prolongado

7.	**strangle**: He strangled her.	
✓	**a.**	**la mató oprimiéndole el cuello**
	b.	le dió todo lo que ella deseaba
	c.	se la llevó por la fuerza
	d.	le tenía una gran admiración

8.	**cavalier**: He treated her in a cavalier manner.	
✓	**a.**	**sin cuidado**
	b.	con amabilidad
	c.	de forma extraña
	d.	como a una hermana

9.	**malign**: His malign influence is still felt.	
✓	**a.**	**malvada**
	b.	buena
	c.	muy importante
	d.	secreta

10.	**veer**: The car veered.	
✓	**a.**	**giró cambiando repentinamente de dirección**
	b.	se sacudió bruscamente de un lado a otro
	c.	hizo un ruido muy fuerte
	d.	se deslizó de un lado a otro sin que las ruedas giraran

Seventh 1000

1.	**olive**: We bought olives.	
✓	**a.**	**fruto aceitoso**
	b.	flores aromáticas de color rosado o rojo
	c.	prenda masculina que se usa para nadar
	d.	utensilios para quitar la maleza

2.	**quilt**: They made a quilt.	
	a.	declaración que hace alguien distribuyendo sus bienes para después de su muerte
	b.	acuerdo firme
✓	**c.**	**cubrecama grueso y caliente**
	d.	instrumento para escribir hecho de pluma

3.	**stealth**: They did it by stealth.	
	a.	gastando una gran suma de dinero
	b.	hiriendo mucho a alguien para que accediera a sus demandas
✓	**c.**	**actuando con el mayor sigilo**
	d.	ignorando los problemas a los que se enfrentaban

4.	**shudder**: The boy shuddered.	
	a.	hablaba en voz baja
	b.	casi se caía
✓	**c.**	**se estremeció**
	d.	lo dijo en voz alta

5.	**bristle**: The bristles are too hard.	
	a.	preguntas
✓	**b.**	**pelos cortos y rígidos**
	c.	camas plegables
	d.	parte del calzado que toca el suelo

6.	**bloc**: They have joined this bloc.	
	a.	grupo musical
	b.	banda de ladrones
	c.	grupo pequeño de soldados enviados antes que a los demás
✓	**d.**	**grupo de países con un propósito en común**

7.	**demography**: This book is about demography.	
	a.	estudio de los patrones del uso de la tierra
	b.	estudio del uso de gráficos para mostrar hechos matemáticos
	c.	estudio del movimiento de las aguas
✓	**d.**	**estudio de la población**

8.	**gimmick**: That's a good gimmick.	
	a.	objeto que permite trabajar de pie por encima del nivel del suelo
	b.	artículo pequeño con bolsillos en cuyo interior se guarda el dinero
✓	**c.**	**acción u objeto que llama la atención**
	d.	plan ingenioso o truco

9.	**azalea**: This azalea is very pretty.	
✓	**a.**	**árbol pequeño con muchas flores que crecen en manojos**
	b.	material liviano hecho de hilos naturales
	c.	traje largo típico de las mujeres de la India
	d.	concha marina que tiene forma de abanico

10.	**yoghurt**: This yoghurt is disgusting.	
	a.	lodo gris oscuro que se haya en el fondo de los ríos
	b.	herida abierta que ha sanado poco
✓	**c.**	**leche agria y espesa que con frecuencia lleva azúcar y saborizante**
	d.	fruto grande, de color púrpura y carne suave

Eighth 1000

1.	**erratic**: He was erratic.	
	a.	libre de culpa
	b.	muy malo
	c.	muy amable
✓	**d.**	**inconstante**

2.	**palette**: He lost his palette.	
	a.	cesta para cargar pescado
	b.	deseo de comer
	c.	joven compañera
✓	**d.**	**tabla en la que el artista mezcla las pinturas**

3.	**null**: His influence was null.	
	a.	tuvo buenos resultados
	b.	no fue fuerte
✓	**c.**	**no tuvo ningún efecto**
	d.	fue duradera

4.	**kindergarten**: This is a good kindergarten.	
	a.	actividad que permite olvidar las preocupaciones
✓	**b.**	**lugar para el aprendizaje de los niños que son muy pequeños para ir a la escuela**
	c.	bolsa resistente y profunda que se lleva sujetada a la espalda
	d.	local donde se puede hacer préstamo de libros

5.	**eclipse**: There was an eclipse.	
	a.	viento fuerte
	b.	ruido fuerte de algo que golpea el agua
	c.	asesinato de un número grande de personas
✓	**d.**	**planeta que ocultó al sol**

6.	**marrow**: This is the marrow.	
	a.	animal que trae buena suerte a un equipo
✓	**b.**	**parte central blanda de los huesos**
	c.	palanca de control para guiar el avión
	d.	aumento de salario

7.	**locust**: There were hundreds of locusts.	
✓	**a.**	**insectos con alas**
	b.	ayudantes que no reciben pago a cambio
	c.	personas que no comen carne
	d.	flores silvestres de colores vivos

8.	**authentic**: It is authentic.	
✓	**a.**	**real**
	b.	muy ruidoso
	c.	viejo
	d.	desértico

9.	**cabaret**: We saw the cabaret.	
	a.	pintura que cubre toda una pared
✓	**b.**	**espectáculo con música y baile**
	c.	pequeño insecto rastrero
	d.	persona con busto de mujer y parte inferior de pez

10.	**mumble**: He started to mumble.	
	a.	reflexionar profundamente
	b.	temblar sin control
	c.	quedarse rezagado
✓	**d.**	**hablar de forma poco clara**

Ninth 1000

1.	**hallmark**: Does it have a hallmark?	
	a.	sello que muestra la fecha de vencimiento
✓	**b.**	**sello que muestra la calidad**
	c.	sello de aprobación de la familia real
	d.	marca o sello empleado para evitar plagio

2.	**puritan**: He is a puritan.	
	a.	persona a quien le gusta llamar la atención
✓	**b.**	**persona con principios morales estrictos**
	c.	persona que no tiene un hogar fijo
	d.	persona que guarda el dinero y detesta gastarlo

3.	**monologue**: Now he has a monologue.	
	a.	lente que se coloca sobre un ojo para ver mejor
✓	**b.**	**turno para hablar por largo tiempo sin ser interrumpido**
	c.	situación en la que se tiene todo el poder
	d.	imagen creada al unir letras de forma ingeniosa

4.	**weir**: We looked at the weir.	
	a.	persona de comportamiento extraño
	b.	terreno pantanoso con plantas acuáticas
	c.	instrumento musical antiguo de viento metal
✓	**d.**	**construcción a lo largo de un río para contener las aguas**

5.	**whim**: He had lots of whims.	
	a.	monedas de oro antiguas
	b.	hembras del caballo
✓	**c.**	**ideas extrañas y sin razón de ser**
	d.	protuberancias enrojecidas y dolorosas

6.	**perturb**: I was perturbed.	
	a.	me hicieron aceptar
✓	**b.**	**preocupado**
	c.	muy aburrido
	d.	muy mojado

7.	**regent**: They chose a regent.	
	a.	persona irresponsable
	b.	persona que dirija la reunión por poco tiempo
✓	**c.**	**persona que actúe en lugar del monarca**
	d.	persona que los represente

8.	**octopus**: They saw an octopus.	
	a.	ave grande que caza de noche
	b.	embarcación para navegar bajo el agua
	c.	aeronave que funciona con hélices giratorias
✓	**d.**	**criatura marina de ocho brazos**

9.	**fen**: The story is set in the fens.	
✓	**a.**	**extensión de tierras bajas y planas cubiertas en parte por agua**
	b.	extensión de tierras altas y montañosas con pocos árboles
	c.	conjunto de casas precarias en una ciudad
	d.	hace mucho tiempo atrás

10.	**lintel**: He painted the lintel.	
✓	**a.**	**viga sobre el borde superior de una puerta o ventana**
	b.	bote para hacer transbordo de un barco grande a tierra
	c.	hermoso árbol de ramas extensas y frutos verdes
	d.	telón sobre el que se proyectan las escenas en el teatro

Tenth 1000

1.	**awe**: They looked at the mountain with awe.	
	a.	preocupación
	b.	interés
✓	**c.**	**asombro**
	d.	respeto

2.	**peasantry**: He did a lot for the peasantry.	
	a.	gente del lugar
	b.	lugar de alabanza
	c.	club de hombres de negocios
✓	**d.**	**agricultores pobres**

3.	**egalitarian**: This organization is very egalitarian.	
	a.	no da mucha información al público
	b.	le disgusta el cambio
	c.	con frecuencia solicita juicios ante el tribunal de justicia
✓	**d.**	**trata a todos sus empleados como iguales**

4.	**mystique**: He has lost his mystique.	
	a.	cuerpo sano
✓	**b.**	**forma secreta para hacer pensar que tenía poderes o destrezas especiales**
	c.	mujer que fue su amante mientras estuvo casado con otra persona
	d.	pelo que crece sobre el labio superior

5.	**upbeat**: I'm feeling really upbeat about it.	
	a.	molesto
✓	**b.**	**bien**
	c.	herido
	d.	confundido

6.	**cranny**: We found it in the cranny!	
	a.	venta de objetos usados
✓	**b.**	**abertura estrecha**
	c.	espacio para guardar cosas debajo del tejado de la casa
	d.	caja grande de madera

7.	**pigtail**: Does she have a pigtail?	
✓	**a.**	**peinado que se realiza trenzando mechones largos de cabello**
	b.	porción larga de tela que cuelga detrás de un vestido
	c.	planta de flores rosadas que cuelgan en ramilletes cortos
	d.	amante

8.	**crowbar**: He used a crowbar.	
✓	**a.**	**barra pesada de metal con un extremo curvo**
	b.	nombre falso
	c.	herramienta afilada usada para abrir agujeros en el cuero
	d.	bastón liviano de metal

9.	**ruck**: He got hurt in the ruck.	
	a.	cavidad situada entre el estómago y la parte superior de las piernas
	b.	riña callejera
✓	**c.**	**en algunos deportes, grupo de jugadores agrupados alrededor de la pelota**
	d.	carrera a lo largo de un campo de nieve

10.	**lectern**: He stood at the lectern.	
✓	**a.**	**mueble que sostiene los libros para leer con comodidad**
	b.	mesa o bloque que se emplea para celebrar el sacrificio de la misa
	c.	lugar donde se consumen bebidas alcohólicas
	d.	borde de la cornisa

Eleventh 1000

1.	**excrete**: This was excreted recently.	
✓	**a.**	**expulsado**
	b.	aclarado
	c.	descubierto mediante un experimento científico
	d.	puesto en una lista de cosas ilegales

2.	**mussel**: They bought mussels.	
	a.	bolitas de vidrio para jugar
✓	**b.**	**mariscos**
	c.	frutas grandes de color púrpura
	d.	piezas de papel suave para mantener la ropa limpia mientras se come

3.	**yoga**: She has started yoga.	
	a.	tejido de hilos hecho a mano
✓	**b.**	**forma de ejercicio para la mente y el cuerpo**
	c.	juego en el que dos jugadores golpean un corcho con plumas
	d.	tipo de danza de los países orientales

4.	**counterclaim**: They made a counterclaim.	
✓	**a.**	**demanda legal hecha por una de las partes para responder a la demanda de la otra parte**
	b.	petición hecha a una tienda para devolver objetos defectuosos
	c.	acuerdo entre dos compañías para intercambiar trabajo
	d.	cobertor para la parte superior de la cama

5.	**puma**: They saw a puma.	
	a.	casa pequeña hecha de adobe
	b.	árbol propio de países calientes y secos
	c.	viento muy fuerte que arrastra todo a su paso
✓	**d.**	**felino salvaje y grande**

6.	**pallor**: His pallor caused them concern.	
	a.	temperatura inusualmente alta
	b.	falta de interés por todo
	c.	grupo de amigos
✓	**d.**	**palidez de la piel**

7.	**aperitif**: She had an aperitif.	
	a.	silla con respaldo largo y un solo posa brazo
	b.	profesor privado de canto
	c.	sombrero grande con plumas largas
✓	**d.**	**bebida que se toma antes de la comida**

8.	**hutch**: Please clean the hutch.	
	a.	rejilla metálica que evita la entrada de la suciedad a las cañerías
	b.	parte trasera de un vehículo usada para guardar bolsas, etc
	c.	eje metálico en el centro de la rueda de las bicicletas
✓	**d.**	**jaula para animales pequeños**

9.	**emir**: We saw the emir.	
	a.	ave con dos plumas largas y curvas en la cola
	b.	mujer que cuida hijos de otros en los países orientales
✓	**c.**	**jefe del Medio Oriente con poder sobre su propio territorio**
	d.	vivienda hecha con bloques de hielo

10.	**hessian**: She bought some hessian.	
	a.	pescado de carne grasa y rosada
	b.	sustancia que produce un estado mental de felicidad
✓	**c.**	**tejido áspero**
	d.	raíz de gusto fuerte usada para sazonar la comida

Twelfth 1000

1.	**haze**: We looked through the haze.	
	a.	ventana pequeña y redonda de los barcos
✓	**b.**	**aire turbio**
	c.	cubierta de láminas de plástico o madera que se ponen en las ventanas
	d.	lista de nombres

2.	**spleen**: His spleen was damaged.	
	a.	hueso de la rodilla
✓	**b.**	**órgano ubicado cerca del estómago**
	c.	tubería que expulsa las aguas residuales de una residencia
	d.	respeto por si mismo

3.	**soliloquy**: That was an excellent soliloquy!	
	a.	canción para seis personas
	b.	dicho corto e ingenioso con un significado profundo
	c.	espectáculo acompañado de luces y música
✓	**d.**	**discurso de un personaje que actúa solo**

4.	**reptile**: She looked at the reptile.	
	a.	libro antiguo escrito a mano
✓	**b.**	**animal de sangre fría y piel cubierta de escamas**
	c.	persona que realiza ventas a domicilio
	d.	técnica artística que consiste en pegar piezas pequeñas de colores diversos

5.	**alum**: This contains alum.	
	a.	sustancia venenosa proveniente de una planta común
	b.	tela suave hecha de hilos artificiales
	c.	tabaco en polvo que se aspira por la nariz
✓	**d.**	**compuesto químico que contiene aluminio**

6.	**refectory**: We met in the refectory.	
✓	**a.**	**salón para comer**
	b.	oficina donde se pueden firmar documentos legales
	c.	espacio en el que pueden dormir varias personas
	d.	espacio con paredes de vidrio para cultivar plantas

7.	**caffeine**: This contains a lot of caffeine.	
	a.	sustancia que da sueño
	b.	filamentos de hojas muy duras
	c.	ideas incorrectas
✓	**d.**	**sustancia estimulante**

8.	**impale**: He nearly got impaled.	
	a.	acusado de un delito grave
	b.	enviado a prisión
✓	**c.**	**atravesado por un objeto puntiagudo**
	d.	involucrado en una disputa

9.	**coven**: She is the leader of a coven.	
	a.	pequeño grupo de canto
	b.	negocio que es propiedad de los trabajadores
✓	**c.**	**sociedad secreta**
	d.	grupo de mujeres que llevan una vida religiosa estricta

10.	**trill**: He practised the trill.	
✓	**a.**	**adorno empleado en una composición musical**
	b.	tipo de instrumento de cuerdas
	c.	forma de lanzar la pelota
	d.	paso de danza que consiste en girar rápido sobre la punta de los dedos de los pies

Thirteenth 1000

1.	**ubiquitous**: Many weeds are ubiquitous.	
	a.	son difíciles de eliminar
	b.	tienen raíces fuertes y largas
✓	**c.**	**se encuentran por todas partes**
	d.	mueren en el invierno

2.	**talon:** Just look at those talons!	
	a.	picos de las montañas
✓	**b.**	**garras afiladas de las aves que cazan**
	c.	capas pesadas de metal para protegerse de las armas
	d.	personas que hacen el ridículo sin darse cuenta

3.	**rouble**: He had a lot of roubles.	
	a.	piedras preciosas de color rojo
	b.	miembros lejanos de la familia
✓	**c.**	**monedas de Rusia**
	d.	limitaciones o dificultades morales

4.	**jovial**: He was very jovial.	
	a.	de muy baja clase social
	b.	muy dado a criticar a otros
✓	**c.**	**muy divertido**
	d.	de muchos amigos

5.	**communiqué**: I saw their communiqué.	
	a.	informe crítico de la organización
	b.	jardín que pertenece a muchos miembros de la comunidad
	c.	material impreso usado con fines publicitarios
✓	d.	**pronunciamiento oficial**

6.	**plankton**: We saw a lot of plankton.	
	a.	hierbas venenosas que se reproducen con gran rapidez
✓	**b.**	**plantas o animales microscópicos que viven en el agua**
	c.	árboles de los que se obtiene madera dura
	d.	arcilla gris que con frecuencia ocasiona derrumbes

7.	**skylark**: We watched a skylark.	
	a.	exhibición de aviones que hacen piruetas en el aire
	b.	objeto, hecho por el hombre, que gira alrededor de la tierra
	c.	persona que hace trucos graciosos
✓	**d.**	**pájaro que vuela alto mientras canta**

8.	**beagle**: He owns two beagles.	
	a.	automóviles rápidos con techos convertibles
	b.	armas grandes que pueden dispararle a mucha gente en poco tiempo
✓	**c.**	**perros pequeños de orejas largas**
	d.	casas construidas para ir de vacaciones

9.	**atoll**: The atoll was beautiful.	
✓	**a.**	**isla coralina de poca elevación que rodea un lago marino**
	b.	obra de arte elaborada tejiendo dibujos con hilos finos
	c.	corona pequeña adornada con muchas joyas preciosas y usadas por las mujeres en la noche
	d.	espacio estrecho y con grandes rocas por el que fluye un río

10.	**didactic**: The story is very didactic	
✓	**a.**	**trata de enseñar algo**
	b.	es muy difícil de creer
	c.	trata de hechos emocionantes
	d.	está escrita de forma poco clara

Fourteenth 1000

1.	**canonical**: These are canonical examples.	
	a.	ejemplos que no siguen las reglas convencionales
	b.	ejemplos tomados de un libro religioso
✓	**c.**	**ejemplos comúnes y ampliamente aceptados**
	d.	ejemplos descubiertos muy recientemente

2.	**atop**: He was atop the hill.	
	a.	en la parte inferior de
✓	**b.**	**en la cima de**
	c.	en la ladera de
	d.	en la parte más alejada de

3.	**marsupial**: It is a marsupial.	
	a.	animal con patas duras
	b.	planta que crece por varios años
	c.	planta con flores que se orientan en dirección al sol
✓	**d.**	**animal provisto de bolsa abdominal para llevar a sus crías**

4.	**augur**: It augured well.	
✓	**a.**	**prometió cosas muy buenas para el futuro**
	b.	estuvo muy de acuerdo con lo que se esperaba
	c.	tenía un color que combinaba bien
	d.	repicó con un sonido claro y hermoso

5.	**bawdy**: It was very bawdy.	
	a.	impredecible
	b.	divertido
	c.	apresurado
✓	**d.**	**grosero**

6.	**gauche**: He was gauche.	
	a.	conversador
	b.	flexible
✓	**c.**	**cohibido**
	d.	decidido

7.	**thesaurus**: She used a thesaurus.	
✓	**a.**	**tipo de diccionario**
	b.	compuesto químico
	c.	manera especial de hablar
	d.	inyección subcutánea

8.	**erythrocyte**: It is an erythrocyte.	
	a.	medicina para calmar el dolor
✓	**b.**	**glóbulo rojo de la sangre**
	c.	metal blanco y rojizo
	d.	miembro de la familia de la ballena

9.	**cordillera**: They were stopped by the cordillera.	
	a.	ley especial
	b.	barco de guerra
✓	**c.**	**cadena de montañas**
	d.	hijo mayor del rey

10.	**limpid**: He looked into her limpid eyes.	
✓	**a.**	**claros**
	b.	llorosos
	c.	marrones
	d.	bellos

Appendix 2: GSL, AWL, Pilot Science List, and new medical word list results over the medical corpus

Appendix 2.1 Coverage and occurrence of the medical corpus by the GSL, AWL, Pilot Science List and the twenty-six medical word lists

WORD LIST	TOKENS #	TOKENS %	TYPES #	TYPES %	FAMILIES #
GSL1	3005823	55.34	4018	7.26	1097
GSL2	352999	6.50	2957	5.34	936
AWL	449900	8.28	2578	4.66	595
Pilot Science List	329236	6.06	1288	2.33	316
MGEN1	330805	6.09	1000	1.81	1000
MED1	279374	5.14	1000	1.81	1000
MGEN2	186676	3.44	1000	1.81	1000
MGEN3	127346	2.34	1000	1.81	1000
MED2	78966	1.45	1000	1.81	1000
MED3	44259	0.81	1000	1.81	1000
MED4	29126	0.54	1000	1.81	1000
MED5	20995	0.39	1000	1.81	1000
MED6	16058	0.30	1000	1.81	1000
MED7	12538	0.23	1000	1.81	1000
MED8	9977	0.18	1000	1.81	1000
MED9	8120	0.15	1000	1.81	1000
MED10	6607	0.12	1000	1.81	1000
MED11	5517	0.10	1000	1.81	1000
MED12	4744	0.09	1000	1.81	1000
MED13	4000	0.07	1000	1.81	1000
MED14	3466	0.06	1000	1.81	1000
MED15	3000	0.06	1000	1.81	1000
MED16	2937	0.05	1000	1.81	1000
MED17	2000	0.04	1000	1.81	1000
MED18	2000	0.04	1000	1.81	1000
MED19	2000	0.04	1000	1.81	1000
MED20	2000	0.04	1000	1.81	1000
MED21	1290	0.02	1000	1.81	1000
MED22	1000	0.02	1000	1.81	1000
MED23	1000	0.02	1000	1.81	1000
Off the lists	107981	1.99	18513	33.44	0
Total	5431740	100	55354	100	28944

Appendix 2.2 The first 1,000 medical word types (content words) occurring both in the medical and general corpora organised by relative frequency

pulmonary	mutations	nodular	exacerbations
anemia	antigen	peptides	splenectomy
biopsy	radiographic	pruritus	shunt
renal	inhibitors	pituitary	epidural
syndromes	hypothyroidism	herpes	epidemiology
receptors	tomography	abnormalities	therapy
lymphoma	angiography	hypertensive	medications
meningitis	intestinal	abdominal	transplantation
venous	optic	biopsies	salivary
fibrosis	capillary	diastolic	lesion
necrosis	coli	titers	diarrhea
infarction	biologic	inhibitor	predispose
distal	cortical	mucous	infections
platelet	dermatitis	maximal	sinuses
autoimmune	antibody	exogenous	congenital
adrenal	cystic	gallbladder	deletion
cirrhosis	infiltrates	endocrine	secreting
interstitial	pathogen	tumors	mutation
receptor	steroid	dysplasia	aorta
serum	adrenergic	mumps	genome
node	postoperative	atrium	afferent
activation	carcinoma	incontinence	antibiotic
invasive	antimicrobial	lactate	prognosis
etiology	mitochondrial	intestine	visceral
biliary	anaerobic	neuron	pancreas
chemotherapy	immunologic	systolic	carotid
sepsis	endogenous	refractory	lobe
intracellular	testosterone	septic	hemorrhages
therapies	lipoprotein	hormone	reactive
aureus	malignancies	metastasis	thoracic
dysfunction	posterior	thyroid	assay
erythema	squamous	catheters	marrow
immunodeficiency	lymphomas	prosthetic	dorsal
plasma	secretions	carcinomas	intravenous
collagen	atrial	arterial	prednisone
toxin	colon	diuretic	obstructive
atrophy	lumen	cranial	nodal
gastric	imaging	hyperthyroidism	ovarian
arrhythmias	metastatic	tetanus	barium
gastrointestinal	inflammatory	metabolites	secrete
peptide	baseline	reductase	nodes
vascular	viral	replication	prostaglandin
ventricular	hepatitis	encoding	rheumatoid
regurgitation	inflammation	bolus	mycobacteria
perfusion	electrolyte	discontinuation	cervical
dialysis	syndrome	hemorrhage	hydrocephalus
metabolic	valvular	toxicity	aseptic
coagulation	anterior	diuretics	femoral
secretion	phosphatase	lymph	placebo
thrombosis	sinus	pathologic	enteric

tumor	gallstones	alveoli	antinuclear
activates	diagnosis	dilated	abscess
prostate	jaundice	inhibition	virulence
malignancy	infarct	obstruction	pathogenic
ventricle	serology	escherichia	granules
dermal	septum	respiratory	bursitis
pharyngeal	mycoplasma	antagonist	tinnitus
proteins	genus	exacerbation	cardiac
infusion	migraine	pathways	axial
nucleotide	radiotherapy	nuclei	duodenal
excretion	occlusion	excreted	hypoplasia
vomiting	hypertension	fetus	leukemias
epilepsy	angina	cartilage	neutralizing
dehydrogenase	glucose	arrhythmia	febrile
viruses	plexus	antagonists	insulin
staining	infection	extrinsic	bacterial
abscesses	disorders	conjugate	enzyme
antibodies	ocular	positron	prophylactic
recurrences	peritonitis	pyridoxine	hormonal
chloroquine	platelets	gamma	gestational
ambulatory	myelin	regimen	leprae
falciparum	nervosa	hypothalamus	metabolism
clonal	penicillin	fructose	degenerative
pathogens	inhibitory	triglycerides	predisposing
leukemia	pelvic	cysts	malignant
chromosome	cyst	overproduction	polypeptide
lipids	abnormality	pigmentation	agar
folic	ulcers	postpartum	contractures
pigmented	bacilli	brainstem	excision
bile	melanoma	botulism	cholecystectomy
diagnostic	toxins	constriction	uremia
bowel	angioplasty	infective	coma
dilation	genitalia	ions	seizures
metabolite	clinician	gallstone	plaques
serotonin	visualization	chromosomes	isolates
solute	somatic	adjuvant	immunologically
lumbar	adipose	encode	sequencing
chronic	steroids	antibiotics	manifestations
duct	electrocardiographic	adhesion	deficiency
morbidity	extensor	clinicians	nucleic
stent	magnesium	tuberculous	legionnaires
lesions	clinical	effusions	gram
cardiovascular	correlate	glaucoma	patients
mellitus	spinal	reticulum	scarring
ulceration	dehydration	impairs	ventilator
clinically	grafts	cell	untreated
genitourinary	neuritis	uric	syphilis
dystrophy	neuroendocrine	encoded	globin
ingestion	pneumonia	lobar	deletions
secreted	testicular	enteritis	neurons

basal	ivermectin	inactivation	physiologically
quadrant	vitro	drowsiness	infects
anesthetic	incubation	helical	histamine
heritable	artery	ligation	biosynthesis
monocytogenes	impairment	enzymes	immunization
dosages	tissue	immunizations	orally
pathway	vectors	interacts	pacemaker
predisposes	noninflammatory	pacemakers	aspirated
peripheral	amantadine	epidermis	diathesis
retinal	intraepithelial	aspirates	activate
rupture	diphtheria	silicosis	proton
liver	behavioral	ligaments	surgical
fractures	situ	repletion	rabies
uterine	intrauterine	follicle	diastole
palsies	ionized	neonates	nonfunctional
fibers	coronary	relapse	symptomatic
nausea	precipitating	symptoms	therapeutic
methionine	endemic	titration	nasal
immunized	cells	intranasal	thymus
potassium	portal	pontine	inhalation
dosage	colic	urea	stimulates
amino	elevations	infectious	bladder
pectoris	clonic	appendicitis	diabetic
clotting	phosphate	genes	bronchial
prevalence	encephalitis	protein	volvulus
skeletal	allergic	degeneration	melanin
substrate	tuberous	encephalomyelitis	dextrose
poliomyelitis	progenitors	vitreous	carbonic
cancers	attenuation	glandular	endoplasmic
cellular	tissues	modality	locus
doses	hospitalized	catheter	regulates
isotonic	disseminated	constipation	ultrasound
nodule	analgesia	ovary	lithium
dosing	tryptophan	utero	dizziness
anorexia	genomes	lactic	fascia
arthritis	overload	matrix	orifice
sclerosis	fetal	asthma	superoxide
neurotransmitters	palliative	tubular	incision
vulgaris	atria	inhibits	carcinogens
transfusion	vibrio	misdiagnosed	vector
fungal	opiates	gouty	gene
sprue	synthetase	monoamine	modalities
plasmodium	gradient	insufficiency	enema
lung	flexion	rheumatic	laxatives
palpitations	implantation	impaired	absorptive
salmonella	synergistic	membranes	vasa
sedimentation	oncogenes	recurrence	triad
thalidomide	dose	abnormal	deposition
optimize	precursors	ulcer	infarctions
stimulatory	normalize	lymphadenitis	amyotrophic

cyanide	intima	fetuses	neurotransmitter
esters	filtration	hydroxylation	antioxidants
dietary	oxalate	molecular	focal
activated	entrapment	prolongation	latex
ion	mastectomy	sedation	genital
analgesics	sclerotic	signaling	fatty
malformations	osteoporosis	hormones	nitric
obesity	cerebral	gland	hypothermia
diabetes	immune	smears	automaticity
genera	lactose	bipolar	kindreds
bronchi	kidney	bronchitis	nicotinamide
elevated	swabs	episodic	disorder
orbital	hookworm	parasites	muscle
recurrent	activating	systole	iodine
microscopy	shunting	quantification	exacerbate
sequestration	infiltration	silica	monoxide
fluoxetine	differential	stigmata	alkali
cramping	gangrene	riboflavin	thrombolytic
ingest	haloperidol	antidepressants	vibratory
hyperinflation	cloning	thrombolysis	menopausal
tuberculosis	macromolecules	parasitic	scan
inactivated	molecules	diagnoses	influenza
induces	infertility	neuralgia	biochemical
suppresses	surgically	discoid	mitotic
tics	organisms	codeine	hydrophobic
trophic	urine	oxidase	technetium
epigenetic	sensory	kidneys	microbe
scans	spleen	fluid	genetic
larynx	brucellosis	bacteria	defect
parasite	inoculated	algorithms	tendon
antioxidant	musculature	precipitates	amplification
neural	yeasts	colonization	gout
diagnosing	virus	leucine	microbiology
disease	microfilariae	mutants	correlates
neonatal	circulating	fibrous	frontal
mortality	diagnosed	defecation	bleeding
membrane	outflow	mimics	stasis
aerobic	aerosols	primates	morphology
transgenic	ulcerated	glycosylation	normalization
protozoa	carbohydrate	delirium	analgesic
nucleated	uptake	vertigo	mites
elucidated	resuscitation	infiltrate	moiety
ductus	transfusions	ingested	ascorbic
sodium	discoloration	tract	catalase
vitamin	interventions	citrate	inactivates
cataracts	dilatation	trauma	puncture
triggers	organism	acids	flukes
transient	bifurcation	urinary	calcified
cerebellum	chromatography	phosphorus	replicate
relapsed	ammonium	emphysema	inhibit

antidepressant	bicarbonate	vials	conjugation
glands	retrograde	curable	larval
therapeutics	lichen	iodinated	erysipelas
globus	encapsulated	aspiration	deflection
mercaptopurine	induced	inhibiting	asthenia
graft	fracture	opportunistic	feedings
outpatient	measles	sleepiness	valve
transplant	soluble	defects	colonized
compression	microscopic	proliferating	allergy
euthanasia	testes	replicates	microorganisms
hoarseness	digits	amphetamine	sickle
obese	glutamate	torsion	ducts
vaccines	dementias	micronutrient	nutrient
larvae	retraction	cognitive	maturation
proteus	inpatient	painless	stimulation
imipramine	periodicity	suppressive	follicles
particulate	pelvis	amphetamines	inducing
islets	screening	asthmatic	arsenic
tibia	microbial	mutant	caffeine
malarial	fractional	hospitalization	inoculum
photon	chills	contaminated	palmar
imperfecta	stimuli	oxidation	circumferential
girdle	mimic	bulimia	oncology
gradients	dissection	swab	metabolizing
depletion	nonhuman	defibrillation	antidotes
ventilation	occult	limbic	transmits
quinine	prostaglandins	paralytic	recovers
avian	ammonia	pyramidal	stabilizes
hemispheric	germinal	synapses	daycare
macroscopic	ferrous	polyethylene	mesylate
mammary	senescence	polypeptides	attenuated
gustatory	lait	transmissible	leprosy
moles	infect	antiemetic	predisposition
spore	nonproductive	stellate	diaphragm
colonizing	circulates	interactions	probes
electrocardiogram	vaccine	mechanisms	anatomically
menopause	calcium	retardation	accelerates
alkaline	crystals	induction	bacterium
molecule	saturation	cramps	granular
trachea	injection	oral	acetate
reflex	cadmium	infected	hereditary
vaginal	hemispheres	acne	fissure
menstrual	median	epidemics	biofeedback
tau	benign	newborns	protrusion
decontamination	bone	synthesis	tsetse
tubercle	mucus	insomnia	quintana
cleaves	remissions	arteries	internists
kinetics	ablative	elevation	splenectomized
anticancer	caudal	inhaled	diseases
relaxants	regenerative	residues	spastic

Appendix 2. 3 The most frequent 1,000 medical word types (content words) occurring only in the medical corpus

systemic	dementia	malabsorption	hematologic
hepatic	thrombocytopenia	vancomycin	neutrophil
neurologic	corticosteroids	glucocorticoid	amyloidosis
myocardial	metastases	radiograph	granulomatous
edema	esophageal	immunosuppressive	neuromuscular
aortic	resection	reabsorption	purpura
asymptomatic	glomerular	intracranial	epithelium
ischemia	bilirubin	syncope	erythematosus
cutaneous	bacteremia	hyperplasia	leukocyte
autosomal	hemolytic	warfarin	antiviral
neuropathy	parenteral	hypertrophy	polymerase
proximal	extracellular	pericardial	aneurysm
endocarditis	neutrophils	aeruginosa	anticoagulation
pancreatic	creatinine	encephalopathy	acyclovir
antigens	angiotensin	monoclonal	cyclosporine
stenosis	histologic	nephropathy	alkalosis
pathogenesis	serologic	ascites	doxycycline
lymphocytes	pneumoniae	endoscopy	apoptosis
acidosis	nodules	albumin	cerebellar
ischemic	sarcoidosis	interferon	nephritis
idiopathic	hypercalcemia	pylori	glomerulonephritis
glucocorticoids	cortisol	cortex	aneurysms
airway	hemorrhagic	neoplasms	hyponatremia
hypotension	cardiomyopathy	proteinuria	pathophysiology
pancreatitis	sputum	parathyroid	dysphagia
prophylaxis	aldosterone	hematopoietic	hypokalemia
vasculitis	endoscopic	subacute	streptococcal
mucosal	autonomic	peritoneal	mononuclear
cytokines	hypersensitivity	myopathy	splenomegaly
hemoglobin	cobalamin	esophagus	fibrillation
heparin	lymphadenopathy	urticaria	synovial
physiologic	assays	scleroderma	osteomyelitis
mitral	ataxia	leukocytes	congestive
regimens	transcription	hyperparathyroidism	eosinophils
tachycardia	anatomic	rifampin	prognostic
estrogen	folate	erythematous	immunosuppression
alveolar	pharmacologic	pericarditis	vasopressin
epithelial	hemolysis	lymphocyte	alleles
macrophages	echocardiography	neutropenia	hypoxemia
lupus	intravascular	fecal	cytotoxic
lipid	subcutaneous	dopamine	emboli
recessive	immunocompromised	lymphocytic	embolism
pleural	immunoglobulin	atherosclerosis	hyperglycemia
endothelial	phenotype	lymphoid	papules
randomized	mol	renin	autoantibodies
dyspnea	myeloma	epidemiologic	reflux
mucosa	androgen	colonic	hypocalcemia
hypoglycemia	colitis	tubule	vertebral
pathobiology	effusion	rectal	pneumococcal

kinase	hypothalamic	hemodynamic	carcinoid
cyclophosphamide	mediastinal	cytoplasmic	adenopathy
candida	follicular	feces	hypovolemia
hypoxia	ceftriaxone	nephrotic	multifocal
radiographs	tricuspid	leishmaniasis	conjunctivitis
contraindicated	antiretroviral	sphincter	hypogonadism
eosinophilia	preexisting	apical	porphyria
cytokine	ampicillin	atherosclerotic	dexamethasone
influenzae	macrophage	neuronal	insipidus
interleukin	globulin	synthase	motility
streptococci	vesicles	parenchymal	immunosuppressed
colorectal	neuropathies	osmolality	bronchiectasis
intravenously	methotrexate	gentamicin	antigenic
noninvasive	catheterization	epidermal	esophagitis
pneumonitis	myocarditis	aplastic	resorption
amyloid	necrotizing	opioid	papillary
adenomas	nonsteroidal	anticoagulant	segmental
medial	ulcerative	subarachnoid	supplementation
percutaneous	protease	peptic	antihistamines
gonadotropin	ejection	musculoskeletal	adenocarcinoma
calcification	cytomegalovirus	pertussis	neoplasm
hyperkalemia	hemodialysis	myocardium	retinopathy
sinusitis	secretory	cellulitis	ribavirin
adenosine	candidiasis	dendritic	hemoptysis
trimethoprim	paraneoplastic	revascularization	corneal
chromosomal	thrombotic	erythromycin	lactam
agonists	septal	prothrombin	anaphylaxis
zoster	pharyngitis	celiac	polymorphonuclear
erythrocyte	phenytoin	allogeneic	necrotic
hepatocellular	titer	anesthesia	colonoscopy
pneumocystis	lymphatic	cholestasis	isoniazid
etiologic	sulfamethoxazole	trachomatis	glycogen
meningococcal	ventilatory	otitis	parenchyma
eosinophilic	erythropoietin	glycoprotein	antibacterial
metronidazole	hemochromatosis	adenoma	contraindications
perforation	postmenopausal	vasoconstriction	distention
monocytes	neoplastic	azithromycin	capillaries
allele	fulminant	psoriasis	cholangitis
tyrosine	bradycardia	ablation	gonococcal
ectopic	sarcoma	azathioprine	paroxysmal
amphotericin	ligand	hypertrophic	cyanosis
hematuria	sequelae	pheochromocytoma	expiratory
simplex	clindamycin	granulomas	tetracycline
varicella	encodes	thrombus	hepatomegaly
medullary	ciprofloxacin	osmotic	polyps
erythrocytes	streptococcus	splenic	humoral
cerebrospinal	myeloid	lamivudine	thyroiditis
leukocytosis	staphylococcus	radiography	myalgias
ultrasonography	fibrin	cerebrovascular	metabolized
nosocomial	mesenteric	erythroid	cephalosporins
aminotransferase	thalassemia	granulomatosis	radiologic

acetylcholine	myasthenia	opioids	testis
hydroxylase	antihypertensive	ganglia	chloramphenicol
immunocompetent	gonadal	cytochrome	histologically
permeability	mastocytosis	gravis	dapsone
thrombin	phenotypic	lavage	activator
relapsing	dyspepsia	overgrowth	enterococci
articular	algorithm	ganciclovir	conjugated
glycemic	histology	neutropenic	intrahepatic
myositis	estrogens	sulfonamides	sarcomas
palpation	cardiopulmonary	granulocyte	dermis
contralateral	polysaccharide	pathophysiologic	shigella
inoculation	cholecystitis	preoperative	uremic
itraconazole	rectum	translocation	urethral
pseudomonas	staphylococcal	comorbid	tubules
gastritis	myopathies	contractility	hematoma
toxoplasmosis	proliferative	autologous	tularemia
neoplasia	amoxicillin	carbamazepine	fibrinogen
rhabdomyolysis	anaerobes	reactivity	legionella
genotype	embolization	intubation	electrolytes
cytoplasm	mineralocorticoid	genomic	aphasia
fibroblasts	demyelinating	effector	antimicrobials
subclinical	angioedema	duodenum	imatinib
vivo	dystonia	mucocutaneous	morphologic
suppressor	hyperuricemia	pleocytosis	triglyceride
chlamydial	osteomalacia	polyneuropathy	thiamine
calcitonin	fistula	vestibular	arteriovenous
arteritis	corticosteroid	androgens	thromboembolism
polycythemia	hematocrit	postural	anticholinergic
anthrax	epinephrine	etiologies	microvascular
parietal	staphylococci	synthesized	gonorrhoeae
gastroenteritis	anion	urate	pyelonephritis
alopecia	acetaminophen	prolapse	myalgia
reactivation	myoclonus	axonal	transferrin
mycobacterial	analogues	pneumothorax	vasculature
apnea	monotherapy	haemophilus	orthostatic
homeostasis	luminal	viremia	methicillin
mycobacterium	cisplatin	sclerosing	endometrial
homozygous	serotypes	amenorrhea	histopathologic
histoplasmosis	urethritis	norepinephrine	glucagon
antifungal	cardiogenic	difficile	ulcerations
aspergillus	oxidative	plasminogen	hypophosphatemia
agonist	gonorrhea	diuresis	ovulation
progesterone	estradiol	leukopenia	polymorphisms
endothelium	hepatotoxicity	campylobacter	aminoglycoside
nystagmus	atrioventricular	digoxin	benzodiazepines
purulent	thyrotoxicosis	heme	enteral
fluconazole	amiodarone	hematogenous	anemias
tamponade	arthralgias	rhinitis	relapses
sulfate	antithrombin	meningeal	retroperitoneal
ipsilateral	fluoroquinolones	somatostatin	infecting

dengue	myeloproliferative	embolic	colchicine
heterozygous	pallidum	prolactin	extrapulmonary
rickets	vasodilation	tricyclic	gondii
basilar	electrophysiologic	statins	pediatric
stricture	occlusive	bullae	nitroglycerin
hyperventilation	purine	chemokines	fibrinolysis
atrophic	hyperthermia	enterovirus	bioavailability
hypoparathyroidism	mediastinum	glomeruli	edematous
nephron	octreotide	monophosphate	acuity
fistulas	adjunctive	opacities	goiter
phosphorylation	clarithromycin	enzymatic	schistosomiasis
transduction	uveitis	ileum	leptin
fluoroquinolone	mononucleosis	erythropoiesis	leptospirosis
inspiratory	airflow	hypoperfusion	parkinsonism
medulla	megaloblastic	multidrug	creatine
echocardiographic	depolarization	phenotypes	phagocytes
loci	hydration	adenovirus	cephalosporin
macular	antiplatelet	ketoacidosis	lymphatics
cytogenetic	clopidogrel	hirsutism	autoimmunity
hyperlipidemia	hyperbilirubinemia	macules	chlamydia
toxicities	polymyositis	quinidine	granuloma
myelopathy	subdural	thiazide	spondylitis
oxygenation	streptomycin	gastrin	claudication
zidovudine	iatrogenic	polymorphic	conjunctival
ferritin	homocysteine	cholinergic	thrombi
antiarrhythmic	nucleoside	polycystic	arginine
dermatomyositis	rotavirus	arterioles	enteroviruses
proteases	petechiae	hepatosplenomegaly	porphyrins
aminoglycosides	empyema	unconjugated	urinalysis
lysis	tacrolimus	radionuclide	hyperpigmentation
perioperative	vesicular	oropharynx	pathognomonic
acromegaly	gadolinium	latency	pericardium
symmetric	tamoxifen	cefotaxime	antiphospholipid
cholestatic	varices	hypernatremia	niacin
hypomagnesemia	anticoagulants	hypoventilation	monocyte
maculopapular	endotracheal	intraabdominal	coagulopathy
steatorrhea	gastroesophageal	intracerebral	bullous
hepatocytes	immunofluorescence	pulmonic	cryoglobulinemia
bactericidal	fibrinolytic	actin	tachyarrhythmias
hyperprolactinemia	transcriptase	allergen	telangiectasia
lipase	catecholamines	retinitis	valacyclovir
hemophilia	chemotherapeutic	phagocytosis	prostatic
oropharyngeal	levodopa	biomarkers	exudate
fasciitis	allopurinol	afterload	nasopharyngeal
aspergillosis	polyuria	diarrheal	mineralization
bronchoscopy	thrombocytopenic	infiltrative	arthropathy
paresthesias	lymphoproliferative	kaposi	neisseria
amylase	parvovirus	thrombocytosis	verapamil
pharynx	noninfectious	stromal	electrophoresis
erosions	fibrotic	progestin	axillary

postprandial	oncogene	galactose	pustules
myosin	inguinal	pestis	alanine
neuropathic	gonadotropins	vasospasm	bioterrorism
trigeminal	hemiparesis	seronegative	phenobarbital
clostridium	imipenem	doxorubicin	adefovir
amebic	pruritic	adenomatous	meningococcemia
mitochondria	lamina	hypertriglyceridemia	neutrophilic
toxoplasma	cofactor	mycophenolate	osteoblasts
thyroxine	polyposis	polyarteritis	photophobia
chorea	echocardiogram	subcutaneously	poliovirus
serotype	bronchiolitis	lipoproteins	porphyrias
hilar	levofloxacin	pleura	pyogenic
hypercalciuria	myopathic	bacillary	tachycardias
azotemia	rituximab	shunts	ureter
dysarthria	psychosis	hypopituitarism	reperfusion
herpesvirus	glycol	midbrain	catabolism
heterozygotes	thalamus	myelitis	keratitis
thromboembolic	demyelination	polyarthritis	prostanoid
theophylline	aspirate	prerenal	virologic
triphosphate	albendazole	stenting	anaphylactic
meningitidis	cryptococcosis	synovitis	lytic
meningoencephalitis	kinases	genotypes	brachial
ligands	pyrimethamine	vasodilators	rashes
carcinoids	steatohepatitis	hybridization	ankylosing
hepatocyte	histocompatibility	anticonvulsants	bacteriuria
pulsatile	prostatitis	chemokine	hypercoagulable
diplopia	bisphosphonates	foramen	interstitium
nitrates	supraventricular	hemoglobinuria	intramuscular
neuroimaging	transmembrane	linezolid	photosensitivity
perianal	stents	trophozoites	spironolactone
unfractionated	allograft	axons	synovium
yersinia	atopic	ileal	resected
reinfection	plantar	neurogenic	giardia
osteoarthritis	infusions	treatable	sagittal
prion	exertional	jejuni	botulinum
mutated	hypoalbuminemia	lactamase	dopaminergic
intracardiac	macrolides	lysosomal	coronavirus
osteoclasts	nephrotoxicity	nodosa	extrahepatic
ptosis	precordial	penicillins	hematopoiesis
pupillary	reticulocyte	aspartate	inotropic
eosinophil	troponin	alkylating	ketoconazole
laryngeal	parasympathetic	hypercholesterolemia	nevi
cava	vasoactive	germline	papilledema
basophils	aplasia	hypovolemic	pentamidine
tinea	bacteroides	macrolide	proinflammatory
constrictive	hyperphosphatemia	preload	contraindication
polyclonal	repolarization	reticular	nasogastric
infarcts	midline	serogroup	hypotonic
immunoglobulins	fibromyalgia	allergens	serine
solutes	urticarial	multifactorial	adrenocorticotropic

Appendix 3: BNC/COCA word list results

Appendix 3.1 Range results over the medical corpus using the twenty-five 1,000 BNC/COCA lists

WORD LIST	TOKENS #	TOKENS %	TYPES #	TYPES %	FAMILIES #
BNC 1st 1000	2822197	51.96	3648	6.59	973
BNC 2nd 1000	634329	11.68	3374	6.10	941
BNC 3rd 1000	568335	10.46	3323	6.00	933
BNC 4th 1000	226094	4.16	2095	3.78	839
BNC 5th 1000	139683	2.57	1670	3.02	764
BNC 6th 1000	93021	1.71	1393	2.52	677
BNC 7th 1000	85163	1.57	1104	1.99	593
BNC 8th 1000	53299	0.98	914	1.65	515
BNC 9th 1000	40536	0.75	721	1.30	451
BNC 10th 1000	39522	0.73	599	1.08	396
BNC 11th 1000	36385	0.67	599	1.08	384
BNC 12th 1000	27908	0.51	448	0.81	314
BNC 13th 1000	33114	0.61	459	0.83	325
BNC 14th 1000	33534	0.62	426	0.77	291
BNC 15th 1000	26482	0.49	415	0.75	288
BNC 16th 1000	22182	0.41	401	0.72	302
BNC 17th 1000	23071	0.42	358	0.65	261
BNC 18th 1000	20308	0.37	364	0.66	265
BNC 19th 1000	15998	0.29	294	0.53	233
BNC 20th 1000	15803	0.29	322	0.58	249
BNC 21st 1000	11619	0.21	271	0.49	223
BNC 22nd 1000	8501	0.16	191	0.35	169
BNC 23rd 1000	9691	0.18	205	0.37	186
BNC 24th 1000	3508	0.06	117	0.21	116
BNC 25th 1000	3279	0.06	156	0.28	124
Off BNC lists	438178	8.07	31487	56.88	0
Total	5431740		55354		10812

Appendix 3.2 Range results over the general corpus using the twenty-five 1,000 BNC/COCA lists

WORD LIST	TOKENS #	TOKENS %	TYPES #	TYPES %	FAMILIES #
BNC 1st1000	4105697	75.59	5924	6.07	999
BNC 2nd1000	464502	8.55	5435	5.57	1000
BNC 3rd1000	265872	4.89	4981	5.10	1000
BNC 4th 1000	100151	1.84	3939	4.03	1000
BNC 5th 1000	58678	1.08	3378	3.46	999
BNC 6th 1000	39106	0.72	3127	3.20	999
BNC 7th 1000	26306	0.48	2701	2.77	997
BNC 8th 1000	20339	0.37	2382	2.44	994
BNC 9th 1000	13964	0.26	2163	2.22	989
BNC 10th 1000	11153	0.21	1929	1.98	977
BNC 11th 1000	8599	0.16	1740	1.78	936
BNC 12th 1000	6236	0.11	1503	1.54	915
BNC 13th 1000	4720	0.09	1210	1.24	829
BNC 14th 1000	3637	0.07	1056	1.08	772
BNC 15th 1000	2503	0.05	882	0.90	660
BNC 16th 1000	2434	0.04	755	0.77	603
BNC 17th 1000	1974	0.04	684	0.70	552
BNC 18th 1000	1280	0.02	531	0.54	450
BNC 19th 1000	1053	0.02	466	0.48	408
BNC 20th 1000	820	0.02	373	0.38	332
BNC 21st 1000	723	0.01	342	0.35	316
BNC 22nd 1000	781	0.01	277	0.28	260
BNC 23rd 1000	753	0.01	222	0.23	208
BNC 24th 1000	769	0.01	198	0.20	192
BNC 25th 1000	372	0.01	146	0.15	139
Off BNC lists	289318	5.33	51304	52.54	0
Total	5431740		97648		17526

Appendix 3.3 Range results over the medical corpus using the twenty-five 1,000 BNC/COCA lists and the BNC/COCA proper noun list

WORD LIST	TOKENS #	TOKENS %	TYPES #	TYPES %	FAMILIES #
BNC 1st1000	2822197	51.96	3648	6.59	973
BNC 2nd1000	634329	11.68	3374	6.10	941
BNC 3rd1000	568335	10.46	3323	6.00	933
BNC 4th 1000	226094	4.16	2095	3.78	839
BNC 5th 1000	139683	2.57	1670	3.02	764
BNC 6th 1000	93021	1.71	1393	2.52	677
BNC 7th 1000	85163	1.57	1104	1.99	593
BNC 8th 1000	53299	0.98	914	1.65	515
BNC 9th 1000	40536	0.75	721	1.30	451
BNC 10th 1000	39522	0.73	599	1.08	396
BNC 11th 1000	36385	0.67	599	1.08	384
BNC 12th 1000	27908	0.51	448	0.81	314
BNC 13th 1000	33114	0.61	459	0.83	325
BNC 14th 1000	33534	0.62	426	0.77	291
BNC 15th 1000	26482	0.49	415	0.75	288
BNC 16th 1000	22182	0.41	401	0.72	302
BNC 17th 1000	23071	0.42	358	0.65	261
BNC 18th 1000	20308	0.37	364	0.66	265
BNC 19th 1000	15998	0.29	294	0.53	233
BNC 20th 1000	15803	0.29	322	0.58	249
BNC 21st 1000	11619	0.21	271	0.49	223
BNC 22nd 1000	8501	0.16	191	0.35	169
BNC 23rd 1000	9691	0.18	205	0.37	186
BNC 24th 1000	3508	0.06	117	0.21	116
BNC 25th 1000	3279	0.06	156	0.28	124
BNC proper noun list	53457	0.98	1885	3.41	1774
Off BNC lists	384721	7.08	29602	53.48	0
Total	5431740		55354		12586

Appendix 3.4 Range results over the general corpus using the twenty-five 1,000 BNC/COCA lists and the BNC/COCA proper noun list

WORD LIST	TOKENS #	TOKENS %	TYPES #	TYPES %	FAMILIES #
BNC 1st1000	4105697	75.59	5924	6.07	999
BNC 2nd1000	464502	8.55	5435	5.57	1000
BNC 3rd1000	265872	4.89	4981	5.10	1000
BNC 4th 1000	100151	1.84	3939	4.03	1000
BNC 5th 1000	58678	1.08	3378	3.46	999
BNC 6th 1000	39106	0.72	3127	3.20	999
BNC 7th 1000	26306	0.48	2701	2.77	997
BNC 8th 1000	20339	0.37	2382	2.44	994
BNC 9th 1000	13964	0.26	2163	2.22	989
BNC 10th 1000	11153	0.21	1929	1.98	977
BNC 11th 1000	8599	0.16	1740	1.78	936
BNC 12th 1000	6236	0.11	1503	1.54	915
BNC 13th 1000	4720	0.09	1210	1.24	829
BNC 14th 1000	3637	0.07	1056	1.08	772
BNC 15th 1000	2503	0.05	882	0.90	660
BNC 16th 1000	2434	0.04	755	0.77	603
BNC 17th 1000	1974	0.04	684	0.70	552
BNC 18th 1000	1280	0.02	531	0.54	450
BNC 19th 1000	1053	0.02	466	0.48	408
BNC 20th 1000	820	0.02	373	0.38	332
BNC 21st 1000	723	0.01	342	0.35	316
BNC 22nd 1000	781	0.01	277	0.28	260
BNC 23rd 1000	753	0.01	222	0.23	208
BNC 24th 1000	769	0.01	198	0.20	192
BNC 25th 1000	372	0.01	146	0.15	139
BNC proper noun list	180997	3.33	13361	13.68	12932
Off BNC lists	108321	1.99	37943	38.86	0
Total	5431740		97648		30458

Appendix 3.5 Range results of the 32,195 medical word types over the twenty-five 1,000 BNC/COCA lists.

WORD LIST	TOKENS #	TOKENS %	TYPES #	TYPES %	FAMILIES #
BNC 1st1000	985	3.06	985	3.06	147
BNC 2nd1000	790	2.45	790	2.45	175
BNC 3rd1000	1017	3.16	1017	3.16	232
BNC 4th 1000	719	2.23	719	2.23	254
BNC 5th 1000	667	2.07	667	2.07	283
BNC 6th 1000	534	1.66	534	1.66	254
BNC 7th 1000	467	1.45	467	1.45	249
BNC 8th 1000	527	1.64	527	1.64	259
BNC 9th 1000	456	1.42	456	1.42	245
BNC 10th 1000	400	1.24	400	1.24	233
BNC 11th 1000	430	1.34	430	1.34	241
BNC 12th 1000	324	1.01	324	1.01	199
BNC 13th 1000	367	1.14	367	1.14	248
BNC 14th 1000	360	1.12	360	1.12	235
BNC 15th 1000	361	1.12	361	1.12	242
BNC 16th 1000	371	1.15	371	1.15	273
BNC 17th 1000	324	1.01	324	1.01	234
BNC 18th 1000	338	1.05	338	1.05	246
BNC 19th 1000	270	0.84	270	0.84	212
BNC 20th 1000	310	0.96	310	0.96	238
BNC 21st 1000	261	0.81	261	0.81	216
BNC 22nd 1000	192	0.60	192	0.60	170
BNC 23rd 1000	198	0.62	198	0.62	180
BNC 24th 1000	112	0.35	112	0.35	111
BNC 25th 1000	144	0.45	144	0.45	116
Off BNC lists	21271	66.07	21271	66.07	0
Total	32195		32195		5492

Appendix 3.6 Number and percentage of medical word families across the BNC/COCA lists

BNC word-family lists	Medical word families	Med%	Medical and general word families
BNC 1st1000	143	14.30	1000
BNC 2nd1000	154	15.40	1000
BNC 3rd1000	205	20.50	1000
BNC 4th 1000	199	19.90	1000
BNC 5th 1000	222	22.20	1000
BNC 6th 1000	186	18.60	999
BNC 7th 1000	194	19.40	1000
BNC 8th 1000	208	20.80	1000
BNC 9th 1000	195	19.50	998
BNC 10th 1000	184	18.60	990
BNC 11th 1000	215	21.90	981
BNC 12th 1000	174	18.10	960
BNC 13th 1000	218	23.60	924
BNC 14th 1000	206	23.30	884
BNC 15th 1000	214	26.10	819
BNC 16th 1000	246	31.70	775
BNC 17th 1000	207	29.00	715
BNC 18th 1000	228	35.10	650
BNC 19th 1000	191	33.00	580
BNC 20th 1000	221	41.20	537
BNC 21st 1000	202	40.60	497
BNC 22nd 1000	162	40.00	405
BNC 23rd 1000	177	47.30	374
BNC 24th 1000	111	37.60	295
BNC 25th 1000	97	38.00	255
Total	4759	24.20	19638
These results show that 37.4% (374 word families) of the 1st 1,000 BNC/COCA word families are medical word families. It also shows that the 23rd 1,000 BNC/COCA list contains the highest percentage of medical headwords (47.3%) in the medical and general corpora.			

Appendix 3.7 Medical word types in the medical and general corpora across the twenty-five 1,000 BNC/COCA lists

BNC word-family lists	Medical word types	Med%	Word types in BNC/COCA lists	Word types in medical and general corpora
BNC 1st 1000	985	14.40	6857	6013
BNC 2nd 1000	790	12.20	6370	5548
BNC 3rd 1000	1017	17.20	5880	5121
BNC 4th 1000	725	14.60	4865	4061
BNC 5th 1000	667	15.50	4294	3511
BNC 6th 1000	535	13.00	4102	3263
BNC 7th 1000	467	12.50	3679	2815
BNC 8th 1000	527	15.00	3419	2534
BNC 9th 1000	456	14.10	3196	2290
BNC 10th 1000	400	13.40	2982	2058
BNC 11th 1000	430	14.50	2942	1935
BNC 12th 1000	324	11.70	2754	1664
BNC 13th 1000	367	15.20	2415	1439
BNC 14th 1000	360	16.00	2299	1301
BNC 15th 1000	361	15.80	2283	1170
BNC 16th 1000	371	17.80	2086	1029
BNC 17th 1000	324	15.60	2076	946
BNC 18th 1000	338	17.50	1933	832
BNC 19th 1000	270	14.40	1872	701
BNC 20th 1000	310	16.80	1820	655
BNC 21st 1000	261	15.70	1651	572
BNC 22nd 1000	192	12.50	1539	446
BNC 23rd 1000	198	14.20	1394	408
BNC 24th 1000	112	8.60	1296	302
BNC 25th 1000	144	8.60	1675	294
Total	10931	14.40	75679	50908

Appendix 3.8 Range results over the medical corpus using the first three 1,000 BNC/COCA lists and twenty seven medical word lists

WORD LIST	TOKENS#	TOKENS%	TYPES#	TYPES%	FAMILIES
BNC 1st 1000	2822197	51.96	3648	6.59	973
BNC 2nd 1000	634329	11.68	3374	6.10	941
BNC 3rd 1000	568335	10.46	3323	6.00	933
MGEN 1st 1000	492781	9.07	1000	1.81	1000
MED 1st 1000	286150	5.27	1000	1.81	1000
MGEN 2nd 1000	129322	2.38	1000	1.81	1000
MED 2nd 1000	81207	1.50	1000	1.81	1000
MGEN 3rd 1000	63382	1.17	1000	1.81	1000
MED 3rd 1000	45478	0.84	1000	1.81	1000
MGEN 4th 1000	36561	0.67	1000	1.81	1000
MED 4th 1000	29880	0.55	1000	1.81	1000
MED 5th 1000	21478	0.40	1000	1.81	1000
MED 6th 1000	16411	0.30	1000	1.81	1000
MED 7th 1000	12807	0.24	1000	1.81	1000
MED 8th 1000	10172	0.19	1000	1.81	1000
MED 9th 1000	8330	0.15	1000	1.81	1000
MED 10th 1000	6716	0.12	1000	1.81	1000
MED 11th 1000	5633	0.10	1000	1.81	1000
MED 12th 1000	4868	0.09	1000	1.81	1000
MED 13th 1000	4000	0.07	1000	1.81	1000
MED 14th 1000	3599	0.07	1000	1.81	1000
MED 15th 1000	3000	0.06	1000	1.81	1000
MED 16th 1000	3000	0.06	1000	1.81	1000
MED 17th 1000	2085	0.04	1000	1.81	1000
MED 18th 1000	2000	0.04	1000	1.81	1000
MED 19th 1000	2000	0.04	1000	1.81	1000
MED 20th 1000	2000	0.04	1000	1.81	1000
MED 21st 1000	1453	0.03	1000	1.81	1000
MED 22nd1000	1000	0.02	1000	1.81	1000
MED 23rd 1000	1000	0.02	1000	1.81	1000
Off BNC MGEN MED lists	130566	2.40	18009	32.53	0
Total	5431740		55354		29847

Appendix 3.9 Range results over the medical corpus using the first four 1,000 BNC/COCA lists

WORD LIST	TOKENS#	TOKENS%	TYPES#	TYPES%	FAMILIES
BNC 1st 1000	2822197	51.96	3648	6.59	973
BNC 2nd 1000	634329	11.68	3374	6.10	941
BNC 3rd 1000	568335	10.46	3323	6.00	933
BNC 4th 1000	226094	4.16	2095	3.78	839
MGEN 1st 1000	389319	7.17	1000	1.81	1000
MED 1st 1000	285365	5.25	1000	1.81	1000
MGEN 2nd 1000	94325	1.74	1000	1.81	1000
MED 2nd 1000	80633	1.48	1000	1.81	1000
MGEN 3rd 1000	53868	0.99	1000	1.81	1000
MED 3rd 1000	45189	0.83	1000	1.81	1000
MED 4th 1000	29709	0.55	1000	1.81	1000
MED 5th 1000	21376	0.39	1000	1.81	1000
MED 6th 1000	16323	0.30	1000	1.81	1000
MED 7th 1000	12739	0.23	1000	1.81	1000
MED 8th 1000	10121	0.19	1000	1.81	1000
MED 9th 1000	8274	0.15	1000	1.81	1000
MED 10th 1000	6686	0.12	1000	1.81	1000
MED 11th 1000	5599	0.10	1000	1.81	1000
MED 12th 1000	4830	0.09	1000	1.81	1000
MED 13th 1000	4000	0.07	1000	1.81	1000
Off BNC MGEN MED lists	112429	2.07	26914	48.62	0
Total	5431740		55354		19686

Appendix 3.10 Range results over the medical corpus including the first five 1,000 BNC/COCA lists

WORD LIST	TOKENS#	TOKENS%	TYPES#	TYPE%	FAMILIES #
BNC 1st 1000	2822197	51.96	3648	6.59	973
BNC 2nd 1000	634329	11.68	3374	6.10	941
BNC 3rd 1000	568335	10.46	3323	6.00	933
BNC 4th 1000	226094	4.16	2095	3.78	839
BNC 5th 1000	139683	2.57	1670	3.02	764
MGEN 1st 1000	315977	5.82	1000	1.81	1000
MED 1st 1000	281796	5.19	1000	1.81	1000
MED 2nd 1000	79744	1.47	1000	1.81	1000
MGEN 2nd 1000	71856	1.32	1000	1.81	1000
MED 3rd 1000	44621	0.82	1000	1.81	1000
MGEN 3rd 1000	32903	0.61	1000	1.81	1000
MED 4th 1000	29328	0.54	1000	1.81	1000
MED 5th 1000	21121	0.39	1000	1.81	1000
MED 6th 1000	16153	0.30	1000	1.81	1000
MGEN 4th 1000	13293	0.24	1000	1.81	1000
MED 7th1000	12606	0.23	1000	1.81	1000
MED 8th 1000	10027	0.18	1000	1.81	1000
Off BNC MGEN MED lists	111677	2.06	29244	52.83	0
Total	5431740		55354		16450

Appendix 3.11 Range results over the medical corpus including the first six 1,000 BNC/COCA lists

WORD LIST	TOKENS#	TOKENS%	TYPES#	TYPE%	FAMILIES #
BNC 1st 1000	2822197	51.96	3648	6.59	973
BNC 2nd 1000	634329	11.68	3374	6.10	941
BNC 3rd 1000	568335	10.46	3323	6.00	933
BNC 4th 1000	226094	4.16	2095	3.78	839
BNC 5th 1000	139683	2.57	1670	3.02	764
BNC 6th 1000	93021	1.71	1393	2.52	677
MED 1st 1000	279978	5.15	1000	1.81	1000
MGEN 1st 1000	257180	4.73	1000	1.81	1000
MED 2nd 1000	78941	1.45	1000	1.81	1000
MGEN 2nd 1000	60022	1.11	1000	1.81	1000
MED 3rd 1000	44266	0.81	1000	1.81	1000
MGEN 3rd 1000	34369	0.63	1000	1.81	1000
MED 4th 1000	29128	0.54	1000	1.81	1000
MED 5th 1000	20993	0.39	1000	1.81	1000
MED 6th 1000	16040	0.30	1000	1.81	1000
MED 7th 1000	12504	0.23	1000	1.81	1000
MED 8th 1000	9949	0.18	1000	1.81	1000
Off BNC MGEN MED lists	104711	1.93	28851	52.12	0
Total	5431740		55354		16127

Appendix 3.12 Range results over the medical corpus including the first nine 1,000 BNC/COCA lists

WORD LIST	TOKENS#	TOKENS%	TYPES#	TYPES%	FAMILIES#
BNC 1st 1000	2822197	51.96	s 3648	6.59	973
BNC 2nd 1000	634329	11.68	3374	6.10	941
BNC 3rd 1000	568335	10.46	3323	6.00	933
BNC 4th 1000	226094	4.16	2095	3.78	839
BNC 5th 1000	139683	2.57	1670	3.02	764
BNC 6th 1000	93021	1.71	1393	2.52	677
BNC 7th 1000	85163	1.57	1104	1.99	593
BNC 8th 1000	53299	0.98	914	1.65	515
BNC 9th 1000	40536	0.75	721	1.30	451
MED 1st 1000	263939	4.86	1000	1.81	1000
MGEN 1st 1000	158433	2.92	1000	1.81	1000
MED 2nd 1000	75996	1.40	1000	1.81	1000
MGEN 2nd 1000	46297	0.85	1000	1.81	1000
MED 3rd 1000	42737	0.79	1000	1.81	1000
MED 4th 1000	28203	0.52	1000	1.81	1000
MED 5th 1000	20429	0.38	1000	1.81	1000
MED 6th 1000	15619	0.29	1000	1.81	1000
MED 7th 1000	12154	0.22	1000	1.81	1000
Off BNC MGEN MED lists	105276	1.94	28112	50.79	0
Total	5431740		55354		15686

Appendix 3.13 Range results over the medical corpus including the first ten 1,000 BNC/COCA lists

WORD LIST	TOKENS#	TOKENS%	TYPES#	TYPE%	FAMILIES #
BNC 1st 1000	2822197	51.96	3648	6.59	973
BNC 2nd 1000	634329	11.68	3374	6.10	941
BNC 3rd 1000	568335	10.46	3323	6.00	933
BNC 4th 1000	226094	4.16	2095	3.78	839
BNC 5th 1000	139683	2.57	1670	3.02	764
BNC 6th 1000	93021	1.71	1393	2.52	677
BNC 7th 1000	85163	1.57	1104	1.99	593
BNC 8th 1000	53299	0.98	914	1.65	515
BNC 9th 1000	40536	0.75	721	1.30	451
BNC 10th 1000	39522	0.73	599	1.08	396
MED 1st 1000	256734	4.73	1000	1.81	1000
MGEN 1st 1000	141222	2.60	1000	1.81	1000
MED 2nd 1000	74918	1.38	1000	1.81	1000
MED 3rd 1000	42227	0.78	1000	1.81	1000
MGEN 2nd 1000	35125	0.65	1000	1.81	1000
MED 4th 1000	27863	0.51	1000	1.81	1000
MED 5th 1000	20231	0.37	1000	1.81	1000
MED 6th 1000	15455	0.28	1000	1.81	1000
MED 7th 1000	12018	0.22	1000	1.81	1000
Off BNC MGEN MED lists	103768	1.91	27513	49.70	0
Total	5431740		55354		16082

Appendix 3.14 Range results over the medical corpus including the twenty-five 1,000 BNC/COCA lists

WORD LIST	TOKENS#	TOKENS%	TYPES#	TYPES%	FAMILIES#
BNC 1st 1000	2822197	51.96	3648	6.59	973
BNC 2nd 1000	634329	11.68	3374	6.10	941
BNC 3rd 1000	568335	10.46	3323	6.00	933
BNC 4th 1000	226094	4.16	2095	3.78	839
BNC 5th 1000	139683	2.57	1670	3.02	764
BNC 6th 1000	93021	1.71	1393	2.52	677
BNC 7th 1000	85163	1.57	1104	1.99	593
BNC 8th 1000	53299	0.98	914	1.65	515
BNC 9th 1000	40536	0.75	721	1.30	451
BNC 10th 1000	39522	0.73	599	1.08	396
BNC 11th 1000	36385	0.67	599	1.08	384
BNC 12th 1000	27908	0.51	448	0.81	314
BNC 13th 1000	33114	0.61	459	0.83	325
BNC 14th 1000	33534	0.62	426	0.77	291
BNC 15th 1000	26482	0.49	415	0.75	288
BNC 16th 1000	22182	0.41	401	0.72	302
BNC 17th 1000	23071	0.42	358	0.65	261
BNC 18th 1000	20308	0.37	364	0.66	265
BNC 19th 1000	15998	0.29	294	0.53	233
BNC 20th 1000	15803	0.29	322	0.58	249
BNC 21st 1000	11619	0.21	271	0.49	223
BNC 22nd 1000	8501	0.16	191	0.35	169
BNC 23rd 1000	9691	0.18	205	0.37	186
BNC 24th 1000	3508	0.06	117	0.21	116
BNC 25th 1000	3279	0.06	156	0.28	124
MED 1st 1000	140451	2.59	1000	1.81	1000
MGEN 1st 1000	101541	1.87	1000	1.81	1000
MED 2nd1000	42027	0.77	1000	1.81	1000
MED 3rd 1000	25331	0.47	1000	1.81	1000
MED 4th 1000	17737	0.33	1000	1.81	1000
MED 5th 1000	13106	0.24	1000	1.81	1000
Off BNC MGEN MED lists	97985	1.80	25487	46.04	0
Total	5431740		55354		16812

Appendix 4: Range results on the med2 corpus

Appendix 4.1 Range results over the med2 corpus using the GSL, AWL and Pilot Science List

Word list	Tokens #	Tokens %	Types #	Types %	Families#
GSL 1st 1132	3280826	55.70	4384	6.75	1110
GSL 2nd 1036	409297	6.95	3365	5.18	982
AWL 600	454065	7.71	2709	4.17	599
Pilot Science List 317	347128	5.89	1332	2.05	316
Off the lists	1399161	23.75	53129	81.84	0
Total	5890477	100.00	64919	100.00	3007

Appendix 4.2 Range results over the med2 corpus using the twenty-five 1,000 BNC/COCA lists

Word list	Tokens#	Tokens%	Types#	Types%	Families
BNC 1st 1000	3114738	52.88	4057	6.25	987
BNC 2nd 1000	678129	11.51	3723	5.73	972
BNC 3rd 1000	570096	9.68	3606	5.55	963
BNC 4th 1000	247054	4.19	2378	3.66	891
BNC 5th 1000	144988	2.46	1901	2.93	841
BNC 6th 1000	99451	1.69	1571	2.42	753
BNC 7th 1000	94464	1.60	1302	2.01	697
BNC 8th 1000	58570	0.99	1074	1.65	590
BNC 9th 1000	44706	0.76	857	1.32	533
BNC 10th 1000	45620	0.77	757	1.17	493
BNC 11th 1000	40038	0.68	707	1.09	446
BNC 12th 1000	32920	0.56	534	0.82	383
BNC 13th 1000	38103	0.65	531	0.82	368
BNC 14th 1000	38098	0.65	485	0.75	333
BNC 15th 1000	28570	0.49	471	0.73	326
BNC 16th 1000	26255	0.45	479	0.74	343
BNC 17th 1000	22832	0.39	410	0.63	290
BNC 18th 1000	21513	0.37	391	0.60	289
BNC 19th 1000	15077	0.26	334	0.51	257
BNC 20th 1000	16732	0.28	365	0.56	278
BNC 21st 1000	11395	0.19	285	0.44	237
BNC 22nd 1000	9133	0.16	220	0.34	194
BNC 23rd 1000	7530	0.13	229	0.35	209
BNC 24th 1000	3201	0.05	143	0.22	142
BNC 25th 1000	3313	0.06	188	0.29	148
Off BNC lists	477951	8.11	37921	58.41	0
Total	5890477	100.00	64919	100.00	11963

Appendix 4.3 Range results over the med2 corpus using the twenty-five BNC/COCA lists and the BNC/COCA proper noun list

Word list	Tokens#	Tokens%	Types#	Types%	Families
BNC 1st 1000	3114738	52.88	4057	6.25	987
BNC 2nd 1000	678129	11.51	3723	5.73	972
BNC 3rd 1000	570096	9.68	3606	5.55	963
BNC 4th 1000	247054	4.19	2378	3.66	891
BNC 5th 1000	144988	2.46	1901	2.93	841
BNC 6th 1000	99451	1.69	1571	2.42	753
BNC 7th 1000	94464	1.60	1302	2.01	697
BNC 8th 1000	58570	0.99	1074	1.65	590
BNC 9th 1000	44706	0.76	857	1.32	533
BNC 10th 1000	45620	0.77	757	1.17	493
BNC 11th 1000	40038	0.68	707	1.09	446
BNC 12th 1000	32920	0.56	534	0.82	383
BNC 13th 1000	38103	0.65	531	0.82	368
BNC 14th 1000	38098	0.65	485	0.75	333
BNC 15th 1000	28570	0.49	471	0.73	326
BNC 16th 1000	26255	0.45	479	0.74	343
BNC 17th 1000	22832	0.39	410	0.63	290
BNC 18th 1000	21513	0.37	391	0.60	289
BNC 19th 1000	15077	0.26	334	0.51	257
BNC 20th 1000	16732	0.28	365	0.56	278
BNC 21st 1000	11395	0.19	285	0.44	237
BNC 22nd 1000	9133	0.16	220	0.34	194
BNC 23rd 1000	7530	0.13	229	0.35	209
BNC 24th 1000	3201	0.05	143	0.22	142
BNC 25th 1000	3313	0.06	188	0.29	148
BNC proper noun list	56427	0.96	2452	3.78	2303
Off BNC lists	421524	7.16	35469	54.64	0
Total	5890477	100.00	64919	100.00	14266

Appendix 4.4 Range results over the med2 corpus using the existing lists (GLS1, GSL2, AWL, Pilot Science List and medical word lists)

Word list	Tokens#	Tokens%	Types#	Types%	Families
GSL1	3280826	55.70	4384	6.75	1110
GSL2	409297	6.95	3365	5.18	982
AWL 600	454065	7.71	2709	4.17	599
SL 317	347128	5.89	1332	2.05	316
MGEN 1st 1000	337567	5.73	996	1.53	996
MED 1st 1000	243919	4.14	993	1.53	993
MGEN 2nd 1000	198036	3.36	994	1.53	994
MGEN 3rd 1000	132741	2.25	977	1.50	977
MED 2nd 1000	72876	1.24	984	1.52	984
MED 3rd 1000	42524	0.72	986	1.52	986
MED 4th 1000	28348	0.48	976	1.50	976
MED 5th 1000	20361	0.35	963	1.48	963
MED 6th 1000	15357	0.26	951	1.46	951
MED 7th 1000	13336	0.23	935	1.44	935
MED 8th 1000	9525	0.16	888	1.37	888
MED 9th 1000	8290	0.14	880	1.36	880
MED 10th 1000	6737	0.11	846	1.30	846
MED 11th 1000	5801	0.10	825	1.27	825
MED 12th 1000	5124	0.09	779	1.20	779
MED 13th 1000	4190	0.07	719	1.11	719
MED 14th 1000	3955	0.07	707	1.09	707
MED 15th 1000	3241	0.06	657	1.01	657
MED 16th 1000	3014	0.05	657	1.01	657
MED 17th 1000	2632	0.04	590	0.91	590
MED 18th 1000	2006	0.03	513	0.79	513
MED 19th 1000	1795	0.03	506	0.78	506
MED 20th 1000	2185	0.04	523	0.81	523
MED 21st 1000	2521	0.04	502	0.77	502
MED 22nd 1000	1039	0.02	352	0.54	352
MED 23rd 1000	1353	0.02	367	0.57	367
Off the lists	230688	3.92	33063	50.93	0
Total	5890477	100.00	64919	100.00	23073

Appendix 4.5 Range results over the med1 corpus using the existing lists (GLS1, GSL2, AWL, Pilot Science List and the new twenty six medical word lists

Word list	Tokens #	Tokens %	Types #	Types %	Families
GSL1	3005823	55.34	4018	7.26	1097
GSL2	352999	6.50	2957	5.34	936
AWL 600	449900	8.28	2578	4.66	595
SL 317	329236	6.06	1288	2.33	316
MGEN 1st 1000	330805	6.09	1000	1.81	1000
MED 1st 1000	279374	5.14	1000	1.81	1000
MGEN 2nd 1000	186676	3.44	1000	1.81	1000
MGEN 3rd 1000	127346	2.34	1000	1.81	1000
MED 2nd 1000	78966	1.45	1000	1.81	1000
MED 3rd 1000	44259	0.81	1000	1.81	1000
MED 4th 1000	29126	0.54	1000	1.81	1000
MED 5th 1000	20995	0.39	1000	1.81	1000
MED 6th 1000	16058	0.30	1000	1.81	1000
MED 7th 1000	12538	0.23	1000	1.81	1000
MED 8th 1000	9977	0.18	1000	1.81	1000
MED 9th 1000	8120	0.15	1000	1.81	1000
MED 10th 1000	6607	0.12	1000	1.81	1000
MED 11th 1000	5517	0.10	1000	1.81	1000
MED 12th 1000	4744	0.09	1000	1.81	1000
MED 13th 1000	4000	0.07	1000	1.81	1000
MED 14th 1000	3466	0.06	1000	1.81	1000
MED 15th 1000	3000	0.06	1000	1.81	1000
MED 16th 1000	2937	0.05	1000	1.81	1000
MED 17th 1000	2000	0.04	1000	1.81	1000
MED 18th 1000	2000	0.04	1000	1.81	1000
MED 19th 1000	2000	0.04	1000	1.81	1000
MED 20th 1000	2000	0.04	1000	1.81	1000
MED 21st 1000	1290	0.02	1000	1.81	1000
MED 22nd 1000	1000	0.02	1000	1.81	1000
MED 23rd 1000	1000	0.02	1000	1.81	1000
Off the lists	107981	1.99	18513	33.44	0
Total	5431740	100.00	55354	100.00	28944

Appendix 4.6 Range results over the med2 corpus using the first three 1,000 BNC/COCA lists and twenty-seven medical word lists

Word list	Tokens#	Tokens%	Types#	Types%	Families
BNC 1st 1000	3114738	52.88	4057	6.25	987
BNC 2nd 1000	678129	11.51	3723	5.73	972
BNC 3rd 1000	570096	9.68	3606	5.55	963
MGEN 1st 1000	500260	8.49	997	1.54	997
MED 1st 1000	248788	4.22	993	1.53	993
MGEN 2nd 1000	138672	2.35	985	1.52	985
MED 2nd 1000	75052	1.27	987	1.52	987
MGEN 3rd 1000	73268	1.24	961	1.48	961
MGEN 4th 1000	52090	0.88	977	1.50	977
MED 3rd 1000	43203	0.73	984	1.52	984
MED 4th 1000	29133	0.49	978	1.51	978
MED 5th 1000	20912	0.36	961	1.48	961
MED 6th 1000	15673	0.27	954	1.47	954
MED 7th 1000	13412	0.23	936	1.44	936
MED 8th 1000	10146	0.17	902	1.39	902
MED 9th 1000	8206	0.14	872	1.34	872
MED 10th 1000	6905	0.12	853	1.31	853
MED 11th 1000	6107	0.10	830	1.28	830
MED 12th 1000	5092	0.09	779	1.20	779
MED 13th 1000	4378	0.07	739	1.14	739
MED 14th 1000	3975	0.07	712	1.10	712
MED 15th 1000	3381	0.06	653	1.01	653
MED 16th 1000	3015	0.05	666	1.03	666
MED 17th 1000	2837	0.05	596	0.92	596
MED 18th 1000	2039	0.03	516	0.79	516
MED 19th 1000	1684	0.03	504	0.78	504
MED 20th 1000	2191	0.04	528	0.81	528
MED 21st 1000	2723	0.05	523	0.81	523
MED 22nd1000	1070	0.02	372	0.57	372
MED 23rd 1000	1370	0.02	354	0.55	354
Off BNC MGEN lists	251932	4.28	32421	49.94	0
Total	5890477	100.00	64919	100.00	24034

Appendix 4.7 Range results over the med2 corpus using the first four 1,000 BNC/COCA lists and fifteen medical word lists

Word list	Tokens#	Tokens%	types#	Types%	Families
BNC 1st 1000	3114738	52.88	4057	6.25	987
BNC 2nd 1000	678129	11.51	3723	5.73	972
BNC 3rd 1000	570096	9.68	3606	5.55	963
BNC 4th 1000	247054	4.19	2378	3.66	891
MGEN 1st 1000	392960	6.67	996	1.53	996
MED 1st 1000	248458	4.22	993	1.53	993
MGEN 2nd 1000	102255	1.74	980	1.51	980
MED 2nd 1000	74337	1.26	987	1.52	987
MGEN 3rd 1000	69679	1.18	979	1.51	979
MED 3rd 1000	43054	0.73	984	1.52	984
MED 4th 1000	28879	0.49	978	1.51	978
MED 5th 1000	20819	0.35	961	1.48	961
MED 6th 1000	15664	0.27	953	1.47	953
MED 7th 1000	13626	0.23	936	1.44	936
MED 8th 1000	9766	0.17	898	1.38	898
MED 9th 1000	8247	0.14	876	1.35	876
MED 10th 1000	6818	0.12	848	1.31	848
MED 11th 1000	5977	0.10	826	1.27	826
MED 12th 1000	5126	0.09	782	1.20	782
MED 13th 1000	4319	0.07	732	1.13	732
Off BNC MGEN lists	230476	3.91	36446	56.14	0
Total	5890477	100.00	64919	100.00	18522

Appendix 4.8 Range results over the med2 corpus using the first five 1,000 BNC/COCA lists and ten medical word lists

Word list	Tokens#	Tokens%	Types#	Type%	Families
BNC 1st 1000	3114738	52.88	4057	6.25	987
BNC 2nd 1000	678129	11.51	3723	5.73	972
BNC 3rd 1000	570096	9.68	3606	5.55	963
BNC 4th 1000	247054	4.19	2378	3.66	891
BNC 5th 1000	144988	2.46	1901	2.93	841
MGEN 1st 1000	320153	5.44	994	1.53	994
MED 1st 1000	244959	4.16	993	1.53	993
MGEN 2nd 1000	76348	1.30	975	1.50	975
MED 2nd 1000	73250	1.24	986	1.52	986
MGEN 3rd 1000	46652	0.79	895	1.38	895
MED 3rd 1000	42475	0.72	986	1.52	986
MED 4th 1000	28634	0.49	976	1.50	976
MGEN 4th 1000	21034	0.36	762	1.17	762
MED 5th 1000	20679	0.35	963	1.48	963
MED 6th 1000	15518	0.26	951	1.46	951
MED 7th 1000	13328	0.23	936	1.44	936
MED 8th 1000	9617	0.16	893	1.38	893
Off BNC MGEN MED lists	222825	3.78	37944	58.45	0
Total	5890477	100.00	64919	100.00	15964

Appendix 4.9 Range results over the med2 corpus using the first six 1,000 BNC/COCA lists and eleven medical lists

Word list	Tokens#	Tokens%	Types#	Type%	Families
BNC 1st 1000	3114738	52.88	4057	6.25	987
BNC 2nd 1000	678129	11.51	3723	5.73	972
BNC 3rd 1000	570096	9.68	3606	5.55	963
BNC 4th 1000	247054	4.19	2378	3.66	891
BNC 5th 1000	144988	2.46	1901	2.93	841
BNC 6th 1000	99451	1.69	1571	2.42	753
MGEN 1st 1000	264022	4.48	994	1.53	994
MED 1st 1000	65439	1.11	962	1.48	962
MED 2nd 1000	56008	0.95	929	1.43	929
MGEN 2nd 1000	243243	4.13	993	1.53	993
MGEN 3rd 1000	72798	1.24	985	1.52	985
MED 3rd 1000	42536	0.72	987	1.52	987
MED 4th 1000	28340	0.48	976	1.50	976
MED 5th 1000	20362	0.35	963	1.48	963
MED 6th 1000	15369	0.26	950	1.46	950
MED 7th 1000	13159	0.22	932	1.44	932
MED 8th 1000	9567	0.16	890	1.37	890
Off BNC MGEN MED lists	205178	3.48	37122	57.18	0
Total	5890477	100.00	64919	100.00	15968

Appendix 4.10 Range results over the med2 corpus using the first nine 1,000 BNC/COCA lists and nine medical word lists

Word list	Tokens#	Tokens%	Types#	Types%	Families
BNC 1st 1000	3114738	52.88	4057	6.25	987
BNC 2nd 1000	678129	11.51	3723	5.73	972
BNC 3rd 1000	570096	9.68	3606	5.55	963
BNC 4th 1000	247054	4.19	2378	3.66	891
BNC 5th 1000	144988	2.46	1901	2.93	841
BNC 6th 1000	99451	1.69	1571	2.42	753
BNC 7th 1000	94464	1.60	1302	2.01	697
BNC 8th 1000	58570	0.99	1074	1.65	590
BNC 9th 1000	44706	0.76	857	1.32	533
MED 1st 1000	228701	3.88	992	1.53	992
MGEN 1st 1000	159914	2.71	983	1.51	983
MED 2nd 1000	69850	1.19	986	1.52	986
MGEN 2nd 1000	58128	0.99	931	1.43	931
MED 3rd 1000	41164	0.70	986	1.52	986
MED 4th 1000	27228	0.46	972	1.50	972
MED 5th 1000	20028	0.34	960	1.48	960
MED 6th 1000	15070	0.26	953	1.47	953
MED 7th 1000	12642	0.21	926	1.43	926
Off BNC MGEN MED lists	205556	3.49	35761	55.09	0
Total	5890477	100.00	64919	100.00	15916

Appendix 4.11 Range results over the med2 corpus using the first ten 1,000 BNC/COCA lists and nine medical lists

Word list	Tokens#	Tokens%	Types#	Type%	Families
BNC 1st 1000	3114738	52.88	4057	6.25	987
BNC 2nd 1000	678129	11.51	3723	5.73	972
BNC 3rd 1000	570096	9.68	3606	5.55	963
BNC 4th 1000	247054	4.19	2378	3.66	891
BNC 5th 1000	144988	2.46	1901	2.93	841
BNC 6th 1000	99451	1.69	1571	2.42	753
BNC 7th 1000	94464	1.60	1302	2.01	697
BNC 8th 1000	58570	0.99	1074	1.65	590
BNC 9th 1000	44706	0.76	857	1.32	533
BNC 10th 1000	45620	0.77	757	1.17	493
MED 1st 1000	222857	3.78	992	1.53	992
MGEN 1st 1000	140678	2.39	979	1.51	979
MED 2nd 1000	69141	1.17	986	1.52	986
MGEN 2nd 1000	46627	0.79	884	1.36	884
MED 3rd 1000	40687	0.69	986	1.52	986
MED 4th 1000	26595	0.45	971	1.50	971
MED 5th 1000	19815	0.34	958	1.48	958
MED 6th 1000	14966	0.25	950	1.46	950
MED 7th 1000	12441	0.21	929	1.43	929
Off BNC MGEN MED lists	198854	3.38	35058	54.00	0
Total	5890477	100.00	64919	100.00	16355

Appendix 4.12 Range results over the med2 corpus using the twenty-five 1,000 BNC/COCA lists and six medical lists

Word list	Tokens#	Tokens%	Types#	Types%	Families
BNC 1st 1000	3114738	52.88	4057	6.25	987
BNC 2nd 1000	678129	11.51	3723	5.73	972
BNC 3rd 1000	570096	9.68	3606	5.55	963
BNC 4th 1000	247054	4.19	2378	3.66	891
BNC 5th 1000	144988	2.46	1901	2.93	841
BNC 6th 1000	99451	1.69	1571	2.42	753
BNC 7th 1000	94464	1.60	1302	2.01	697
BNC 8th 1000	58570	0.99	1074	1.65	590
BNC 9th 1000	44706	0.76	857	1.32	533
BNC 10th 1000	45620	0.77	757	1.17	493
BNC 11th 1000	40038	0.68	707	1.09	446
BNC 12th 1000	32920	0.56	534	0.82	383
BNC 13th 1000	38103	0.65	531	0.82	368
BNC 14th 1000	38098	0.65	485	0.75	333
BNC 15th 1000	28570	0.49	471	0.73	326
BNC 16th 1000	26255	0.45	479	0.74	343
BNC 17th 1000	22832	0.39	410	0.63	290
BNC 18th 1000	21513	0.37	391	0.60	289
BNC 19th 1000	15077	0.26	334	0.51	257
BNC 20th 1000	16732	0.28	365	0.56	278
BNC 21st 1000	11395	0.19	285	0.44	237
BNC 22nd 1000	9133	0.16	220	0.34	194
BNC 23rd 1000	7530	0.13	229	0.35	209
BNC 24th 1000	3201	0.05	143	0.22	142
BNC 25th 1000	3313	0.06	188	0.29	148
MED 1st 1000	109263	1.85	981	1.51	981
MGEN 1st 1000	104744	1.78	926	1.43	926
MED 2nd1000	37202	0.63	979	1.51	979
MED 3rd 1000	22935	0.39	957	1.47	957
MED 4th 1000	16507	0.28	952	1.47	952
MED 5th 1000	12589	0.21	928	1.43	928
Off BNC MGEN MED lists	174711	2.97	32198	49.60	0
Total	5890477	100.00	64919	100.00	17686

Appendix 5: Ethics approval

Appendix 5.1 Memorandum issued by Victoria University of Wellington Human Ethics Committee

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

MEMORANDUM

Phone 0-4-463 5676

TO	Betsy Quero
COPY TO	Paul Nation, Averil Coxhead
FROM	Dr Allison Kirkman, Convener, Human Ethics Committee
DATE	31 August 2010
PAGES	1
SUBJECT	**Ethics Approval: No 17912 The vocabulary of academic texts**

Thank you for your applications for ethical approval, which have now been considered by the Standing Committee of the Human Ethics Committee.

Your applications have been approved from the above date and this approval continues until 2 August 2011. If your data collection is not completed by this date you should apply to the Human Ethics Committee for an extension to this approval.

Best wishes with the research.

Allison Kirkman
Human Ethics Committee

Appendix 6: Consent form

Appendix 6.1 Consent form for VST administration

CONSENT TO PARTICIPATE IN RESEARCH

Project: The *vocabulary load of academic texts*

I have been given an explanation of this research and I have understood it.

I have had an opportunity to ask questions and have had them answered to my satisfaction.

I understand that the information I provide will be kept confidential to the research, the published results will not use my name and that no opinions will be attributed to me in any way that will identify me.

I understand that I may withdraw from this research at any time before 31 July, 2011, without having to give reasons. If I withdraw, my data will be destroyed immediately.

I understand that when this research is completed the information obtained will be destroyed after five (5) years.

I understand that, if I so request, details of the purpose of the research will be discussed with me once I have completed my participation.

I understand that I may ask for a summary of results from the study to be sent to me at a later date, and write below an email address to which this summary can be sent.

I agree to take part in this research.

Name: (please print clearly): _____

Signed: _____

Email address: _____

Date: _____

CONTACT DETAILS

<table>
<tr><td colspan="2">Betsy Quero (PhD student)</td></tr>
<tr><td>BETSY'S CONTACT DETAILS IN NEW ZEALAND
School of Linguistics and Applied Language Studies
Victoria University of Wellington
PO BOX 600, Wellington 6140
NEW ZEALAND
Tel: +64 4 463 5233 Extn 8709
Email: Betsy.Quero@vuw.ac.nz</td><td>BETSY'S CONTACT DETAILS IN VENEZUELA
Escuela de Idiomas Modernos
Facultad de Humanidades y Educ
Edificio A, 3er Piso. Sector La Liria
Merida, VENEZUELA
Tel:+582742401880
Fax:+582742401899
Email: Betsy.Quero@vuw.ac.nz</td></tr>
<tr><td>Paul Nation (Supervisor)
School of Linguistics and Applied Language Studies
Victoria University of Wellington
PO BOX 600, Wellington 6140
NEW ZEALAND
Tel: +64 4 463 5628
Email: Paul.Nation@vuw.ac.nz</td><td>Averil Coxhead (Supervisor)
School of Linguistics and Applied Language Studies
Victoria University of Wellington
PO BOX 600, Wellington 6140
NEW ZEALAND
Tel: +64 4 463 5625
Email: Averil.Coxhead@vuw.ac.nz</td></tr>
</table>

Appendix 6.2 Consent form for VST administration (Translated into Spanish)

La versión monolingüe de esta prueba de medición de vocabulario pueden encontrarla en la siguiente dirección: http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

CONSENTIMIENTO A PARTICIPAR EN LA INVESTIGACIÓN

Projecto: *La carga de vocabulario de los textos académicos*

Me explicaron en qué consiste esta investigación y entiendo cuál es su propósito.

He tenido la oportunidad de hacer preguntas y he quedado conforme con las respuestas que he recibido.

Entiendo que la información que me suministraron será guardada de forma confidencial durante la investigación. También entiendo que mi nombre no aparecerá en los resultados que se publiquen de esta investigación y que no será posible relacionar las opiniones que yo emita con mi nombre.

Entiendo que puedo retirarme de esta investigación antes del 31 de julio de 2011, sin tener que explicar las razones de mi retiro. Si decido retirarme de esta investigación, la información que he suministrado será destruida inmediatamente.

Entiendo que cuando esta investigación concluya la información que se obtenga de este examen será destruida a los cinco (5) años de haber culminado este estudio.

Entiendo que, si así lo solicitase, los detalles del propósito de esta investigación serán discutidos conmigo una vez que haya culminado mi participación.

Entiendo que puedo solicitar que posteriormente se me haga llegar un resumen de los resultados de este estudio. Para ello escribo a continuación una dirección de correo electrónico a la que se me puede enviar este resumen.

Estoy de acuerdo en participar en esta investigación.

Nombre y apellido: (use letra clara y legible): _____

Firma: _____

Dirección de correo electrónico: _____

Fecha: _____

DIRECCIONES DE LOS CONTACTOS

Betsy Quero (Estudiante de doctorado)	
DIRECCION DE BETSY EN NUEVA ZELANDA School of Linguistics and Applied Language Studies Victoria University of Wellington PO BOX 600, Wellington 6140 NEW ZEALAND Tel: +64 4 463 5233 Extn 8709 Email: Betsy.Quero@vuw.ac.nz	**DIRECCION DE BETSY EN VENEZUELA** Escuela de Idiomas Modernos Facultad de Humanidades y Educ Edificio A, 3er Piso. Sector La Liria Merida, VENEZUELA Tel:+582742401880 Fax:+582742401899 Email: Betsy.Quero@vuw.ac.nz
Paul Nation (Tutor de tesis) School of Linguistics and Applied Language Studies Victoria University of Wellington PO BOX 600, Wellington 6140 NEW ZEALAND Tel: +64 4 463 5628 Email: Paul.Nation@vuw.ac.nz	Averil Coxhead (Tutora de tesis) School of Linguistics and Applied Language Studies Victoria University of Wellington PO BOX 600, Wellington 6140 NEW ZEALAND Tel: +64 4 463 5625 Email: Averil.Coxhead@vuw.ac.nz

Appendix 7: Information sheet

Appendix 7.1 Information sheet for VST administration

Title of project: The vocabulary load of academic texts

PARTICIPANT INFORMATION SHEET

My name is Betsy Quero and I am a PhD student in the School of Linguistics and Applied Language Studies at Victoria University of Wellington, New Zealand. As a thesis candidate, my research topic is 'The vocabulary load of academic texts'. One of the aspects that this research needs to investigate is the vocabulary size that native Spanish speakers taking English for Specific Purposes (ESP) classes need to meet the vocabulary demands of academic texts written in English. Research on specialised vocabulary has reported the general trend among native speakers of Romance languages (French, Italian, Portuguese and Spanish) to perform well in academic vocabulary tests in English.

During this stage of my research, I need to conduct a vocabulary test to estimate vocabulary size. The results of this vocabulary size test may provide a more realistic picture of the vocabulary load that academic texts in English have for L1 speakers of Romance languages (L1 Spanish speakers, in particular). Since the results of the research are assumed to have implications for ESP learners, that is why you are invited to take part in this study which has received approval from the Human Ethics Committee of Victoria University of Wellington.

This study involves gathering interview data to develop the standard Spanish translation of Nation's (2007) Vocabulary Size Test (VST). This interview data will be gathered emailing the Spanish version of the VST to five speakers of different dialects of Latin American Spanish (Mexican, Argentinean, Chilean, Colombian, and Peruvian) and three speakers of Peninsular Spanish. While you answer the test I would also appreciate your comments on any aspects of the Spanish language you consider worth mentioning, such as word choice and clarity. Likewise, a final comment providing some overall feedback about this test taking situation would also be appreciated.

If any part of the procedure makes you uncomfortable for any reason, you can withdraw from the study at any time before 31 July, 2011, making the request via email.
Your answer sheet includes your name and a number to help me organise the results. The data will be stored in a secure place at all times. The statistically analysed test data will be used for scholarly conference presentation, scholarly journal articles to be published, and the PhD thesis to

be submitted for marking to the school and deposited in the University Library. Answer sheets will be destroyed five years after the end of the project.
If you have any questions or would like to receive further information about this research, please contact me or my supervisors.

Thank you for reading through this information sheet.

CONTACT DETAILS

Betsy Quero (PhD student)	
BETSY'S CONTACT DETAILS IN NEW ZEALAND School of Linguistics and Applied Language Studies Victoria University of Wellington PO BOX 600, Wellington 6140 NEW ZEALAND Tel: +64 4 463 5233 Extn 8709 Email: Betsy.Quero@vuw.ac.nz	**BETSY'S CONTACT DETAILS IN VENEZUELA** Escuela de Idiomas Modernos Facultad de Humanidades y Educacion Edificio A, 3er Piso. Sector La Liria Merida, VENEZUELA Tel:+582742401880 Fax:+582742401899 Email: Betsy.Quero@vuw.ac.nz
Paul Nation (Supervisor) School of Linguistics and Applied Language Studies Victoria University of Wellington PO BOX 600, Wellington 6140 NEW ZEALAND Tel: +64 4 463 5628 Email: Paul.Nation@vuw.ac.nz	Averil Coxhead (Supervisor) School of Linguistics and Applied Language Studies Victoria University of Wellington PO BOX 600, Wellington 6140 NEW ZEALAND Tel: +64 4 463 5625 Email: Averil.Coxhead@vuw.ac.nz

Appendix 7.2 Information sheet for VST administration (Translated into Spanish)

Título del proyecto: La carga del vocabulario de los textos académicos

DOCUMENTO DE INFORMACION SOBRE LA PRUEBA

Mi nombre es Betsy Quero. Soy estudiante de doctorado en la Escuela de Lingüística y Estudios de Lingüística Aplicada de la Universidad Victoria de Wellington, Nueva Zelanda. El tema de investigación de mi tesis es 'La carga de vocabulario de los textos académicos'. Uno de los aspectos estudiados en esta investigación se refiere a la cantidad de vocabulario que los hablantes nativos de español que realizan estudios de Inglés con Propósitos Específicos (IPE) necesitan para cumplir con las exigencias léxicas de los textos académicos escritos en la lengua inglesa. Las investigaciones realizadas sobre vocabulario especializado han mencionado la tendencia general que tienen los hablantes de lenguas romances (francés, italiano, portugués y español) de salir bien en pruebas de medición del vocabulario académico en inglés.

En esta fase de mi investigación, debo aplicar una prueba de vocabulario que me permita determinar el tamaño del vocabulario de hablantes de español que estudian IPE. Espero que los resultados de esta prueba de vocabulario ofrezcan una visión más real de la carga de vocabulario que presentan los textos académicos para los hablantes nativos de las lenguas romances, en particular del español. Debido a que los resultados de esta investigación podrían incidir en los aprendices de IPE, usted ha sido invitado a participar en este estudio. Esta investigación cuenta con la aprobación del Comité de Ética Humana de la Universidad Victoria, de Wellington. Es también importante destacar que su participación en esta investigación no influirá de manera alguna en la calificación final del curso de inglés instrumental.

De los resultados de este estudio piloto, así como de los comentarios y sugerencias que suministre mientras responde la prueba, se espera obtener información que permita unificar criterios y producir la versión en español estándar de esta prueba de medición de vocabulario diseñada en inglés por Nation (2007). Esta información espero obtenerla enviando esta prueba por correo electrónico a cinco hablantes de diferentes dialectos de español latinoamericano (mexicano, argentino, chileno, colombiano, peruano) y tres hablantes nativos de español peninsular. Por lo antes expuesto, le agradecería altamente que mientras responda esta prueba incluyera comentarios sobre cualquier aspecto de la lengua española que considere importante mencionar, como por ejemplo, escogencia del vocabulario, uso de las expresiones y claridad. Asimismo, de ser posible agradecería incluyera un comentario general una vez haya terminado de leer y responder la prueba.

Si algún aspecto de la administración de esta prueba lo incomoda, usted puede expresar su deseo de retirarse de este estudio antes del 31 de julio de 2011. Esta solicitud de retiro de esta prueba de vocabulario puede hacerla vía correo electrónico.
La hoja de respuestas incluye su nombre y un número con el fin de ayudarme a organizar los resultados. Estos resultados serán guardados, todo el tiempo, en un lugar seguro. Los resultados del análisis estadístico de esta prueba podrán usarse para conferencias, exposiciones orales, artículos en revistas especializadas y en la tesis de doctorado la cual será guardada en la

biblioteca de la Universidad Victoria. Las hojas de respuestas de esta prueba serán destruidas a los 5 años después de haber concluido este estudio.

Si desea hacer alguna pregunta o si le gustaría recibir algún tipo de información adicional sobre esta investigación, favor ponerse en contacto conmigo o con mi alguno de mis supervisores.

Gracias por su tiempo y participación.

DIRECCIONES DE LOS CONTACTOS

<table>
<tr><td colspan="2">Betsy Quero (Estudiante de doctorado)</td></tr>
<tr><td>DIRECCION DE BETSY EN NUEVA ZELANDA
School of Linguistics and Applied Language Studies
Victoria University of Wellington
PO BOX 600, Wellington 6140
NEW ZEALAND
Tel: +64 4 463 5233 Extn 8709
Email: Betsy.Quero@vuw.ac.nz</td><td>DIRECCION DE BETSY EN VENEZUELA
Escuela de Idiomas Modernos
Facultad de Humanidades y Educacion
Edificio A, 3er Piso. Sector La Liria
Merida, VENEZUELA
Tel:+582742401880
Fax:+582742401899
Email: Betsy.Quero@vuw.ac.nz</td></tr>
<tr><td>Paul Nation (Tutor de tesis)
School of Linguistics and Applied Language Studies
Victoria University of Wellington
PO BOX 600, Wellington 6140
NEW ZEALAND
Tel: +64 4 463 5628
Email: Paul.Nation@vuw.ac.nz</td><td>Averil Coxhead (Tutora de tesis)
School of Linguistics and Applied Language Studies
Victoria University of Wellington
PO BOX 600, Wellington 6140
NEW ZEALAND
Tel: +64 4 463 5625
Email: Averil.Coxhead@vuw.ac.nz</td></tr>
</table>